AuthentiCity: A Multi-Source Provenance-Aware Knowledge Graph and Benchmark for 3D City Models

Huynh Duc An Son Nguyen
son.nguyen@hcu-hamburg.de
HafenCity University Hamburg,
Computational Methods Lab
Hamburg, Germany

Lukas Arzoumanidis
lukas.arzou@hcu-hamburg.de
HafenCity University Hamburg,
Computational Methods Lab
Hamburg, Germany

Youness Dehbi
youness.dehbi@hcu-hamburg.de
HafenCity University Hamburg,
Computational Methods Lab
Hamburg, Germany

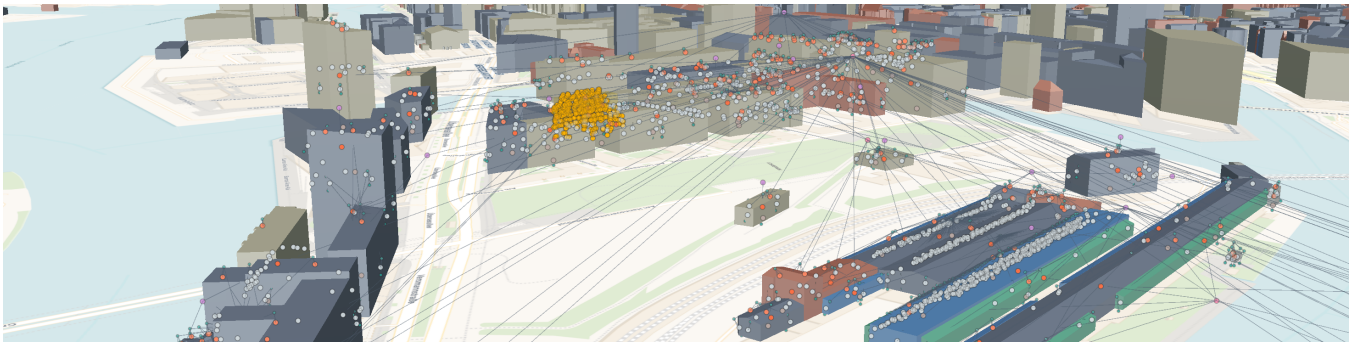


Figure 1: Hamburg, one of five *AuthentiCity* datasets: a provenance-aware 3D knowledge graph integrating CityGML, OpenStreetMap, roof-material predictions (color-coded buildings), and reconstructed LoD3 geometry (orange subgraph).

Abstract

Urban digital twins increasingly combine authoritative, crowdsourced, machine-learned, and reconstructed data with differing reliability, coverage, and semantics. Yet few urban datasets provide a unified representation that supports multi-source integration, provenance tracking, spatial reasoning, and machine learning. As a result, existing benchmarks rarely evaluate reasoning about source origin, confidence, coverage, and agreement. We present *AuthentiCity*, a multi-source, provenance-aware 3D city knowledge graph spanning five cities across three continents (Hamburg, Helsinki, Zurich, New York, and Tokyo) and comprising 180 GiB, 180M nodes, 220M edges, 1.2B properties, and 3.6M buildings. The labeled property graphs integrate authoritative CityGML features with OpenStreetMap data for all cities, adding roof-material predictions and reconstructed LoD3 geometry for Hamburg, under a provenance model in which derived information never replaces authoritative data. Confidence-weighted edges resolve many-to-many cross-source correspondences, constructing canonical urban entities while preserving traceable links to all contributing evidence. *AuthentiCity* is primarily a data contribution. We introduce two benchmark families to demonstrate the tasks enabled by the representation. The first evaluates natural-language-to-query translation with tasks beyond conventional text-to-SQL and text-to-Cypher benchmarks, including 3D spatial reasoning, provenance-aware filtering, cross-source agreement and disagreement, coverage-aware aggregation, and infeasible-query detection. The second evaluates graph representation learning through multi-source attribute prediction, node classification, and cross-source matching prediction, facilitating comparison of provenance-agnostic and provenance-aware embeddings. Even a strong commercial LLM reaches only 54–69 % execution accuracy and a 7B open-weight model 6–19 %, and the open-weight model never abstains on an unanswerable question. We release the complete artifact under open licenses with an archival DOI: the five enriched property graphs, loaders, the question suite with gold queries and materialized answers, task splits, an evaluation harness, and a datasheet.

CCS Concepts

• **Information systems** → **Graph-based database models**.

Keywords

Dataset, Benchmark, CityGML, OSM, Knowledge Graph

1 Introduction

More than half of the world's population lives in cities [55], and national mapping agencies increasingly provide authoritative semantic 3D city models at city, regional, and national scales [3, 29, 44]. Typically encoded in CityGML [24, 32], these models provide virtual representations of urban environments and form a key component of urban digital twins (UDTs) for planning, simulation, and decision-making [31, 34]. Yet no single source provides a complete description of a city. Cadastral models offer surveyed geometry and official semantics but often lack use-level detail. Crowd-sourced maps such as OpenStreetMap (OSM) add names, addresses, points of interest, and street networks, but vary in coverage and quality [6]. Machine learning can infer attributes such as roof materials from orthophotos [4, 30], while photogrammetry provides detailed facade geometry. Integrating these sources therefore requires reasoning over facts that differ in origin, confidence, and coverage.

Existing benchmarks do not directly evaluate this capability. Text-to-query benchmarks such as Spider [65], BIRD [36], and CypherBench [21] focus on relational or encyclopedic data and do not combine spatial predicates, 3D city geometry, and source provenance. Urban benchmarks such as CityBench [20] and UUKG [47] evaluate urban reasoning or spatiotemporal prediction, but neither exposes a queryable semantic 3D city model nor represents differing trust levels across integrated sources. Consequently, current benchmarks cannot test whether systems select authoritative rather than predicted values, qualify uncertain evidence, or recognize unsupported aggregates arising from incomplete coverage. We refer to these capabilities as *provenance-aware reasoning*.

We introduce *AuthentiCity*, a multi-city knowledge graph (KG) and benchmark for evaluating provenance-aware querying and representation learning over heterogeneous 3D city data. The KG is designed around traceability: authoritative, crowd-sourced, predicted, and reconstructed information are represented as distinct entities and relations. External objects remain separate nodes, fusion edges record matching confidence, and source, confidence, and coverage become explicit inputs to benchmark evaluation. *AuthentiCity* is constructed with *pykci* [45], an open-source pipeline for mapping CityGML datasets to a compact labeled property graph (LPG) in Neo4j with an R-tree spatial index. On top of the authoritative layer, *AuthentiCity* integrates OSM through confidence-weighted matching edges, attaches machine-learned roof-material predictions, and incorporates reconstructed LoD3 building models.

The dataset follows a two-tier design that balances cross-city comparability with source depth. Tier 1 applies the same CityGML-plus-OSM construction to all cities, enabling a comparable evaluation core. Tier 2 provides a deep-fusion instance for Hamburg, where all four sources are available: CityGML, OSM, ML predictions, and LoD3 reconstructions. The dataset also preserves incomplete coverage rather than masking it. For example, roof-material predictions are available for only about half of Hamburg's buildings, enabling evaluation of whether models distinguish unavailable facts from negative facts and avoid unsupported city-wide conclusions.

We define two complementary benchmark task families. The first evaluates natural-language-to-query translation with gold Cypher over the property graph, covering spatial predicates, cross-source agreement and disagreement, provenance-filtered retrieval, coverage-aware aggregation, and infeasible-question detection (Section 4). The second evaluates graph representation learning through attribute imputation, node classification, and matching-link prediction, comparing provenance-agnostic and provenance-aware embeddings (Section 5). Together, these tasks evaluate provenance awareness at both the symbolic and representation-learning levels.

Our contributions are:

- **A provenance-preserving, multi-source 3D city dataset.** We release five two-tier city KGs with 180 GiB of data, 180 million nodes, 220 million edges, 1.2 billion properties, and 3.6 million buildings. The KG integrates authoritative CityGML data, crowd-sourced OSM data, ML-predicted attributes, and reconstructed LoD3 geometry while preserving source provenance through canonical feature nodes and source-specific attachments. The release includes loaders, benchmark splits, a datasheet [22], and an archival DOI.
- **A benchmark for provenance-aware querying and graph learning.** We provide a text-to-query suite covering spatial, cross-source, coverage-aware, and infeasibility cases, together with a representation-learning suite for attribute imputation, node classification, and matching-link prediction.
- **Baselines and diagnostic analysis.** We report reference results and a failure decomposition that separates Cypher-generation errors from semantic errors and from failures to abstain. For graph learning, we compare provenance-agnostic and provenance-aware variants to measure the value of explicitly representing source information.

2 Related Work

2.1 Natural-Language-to-Query Benchmarks

Text-to-SQL is the most mature setting. Spider [65] established cross-domain evaluation (10,181 questions over 200 databases) but its schemas are small and carry no spatial types. BIRD [36] scaled to 12,751 questions over 95 large, noisy databases and introduced the execution-centric evaluation we adopt, yet still contains no geospatial reasoning and treats every value as equally trustworthy. Spider 2.0 [35] adds enterprise-scale realism but remains relational. NL2SQL-BUGs [39] shifts focus to detecting semantically incorrect SQL, which motivates our infeasibility category, and Dr.Spider [13] stresses robustness under perturbations; neither addresses spatial reasoning, provenance, or source disagreement.

For property graphs, the Neo4j Text2Cypher dataset [50] aggregates about 44,000 instances, though many are not grounded in an executable graph. CypherBench [21] provides 11 Wikidata-derived graphs with over 10,000 questions and the execution-accuracy metric we adopt, but its graphs are encyclopedic, with no geometry. Mind the Query [14] contributes more than 27,000 validated text-to-Cypher pairs with a rigorous validation pipeline we take as a quality template. All share two limits relevant here: their schemas contain neither spatial geometry nor multiple sources describing the same entity, so the capabilities *AuthentiCity* targets are outside their scope.

2.2 Urban Benchmarks and Knowledge Graphs

CityBench [20] evaluates LLMs and VLMs on eight urban tasks across 13 cities but exposes no queryable semantic 3D city model. CityGPT [19] embeds urban knowledge into the model itself, which does not generalize across cities or updates and gives no auditable grounding. UUKG [47] releases unified urban KGs for spatiotemporal prediction and UrbanKGent [46] automates KG construction with LLM agents, but both operate on POI- and region-level entities rather than 3D building models, and neither provides queryable provenance. Surveys of urban KGs and digital twins [2, 40, 58, 60] consistently name data fusion as a primary motivation, yet provenance is rarely a first-class queryable component.

Closest to our sources, Ding et al. [16] integrate CityGML and OSM into an RDF KG queryable with GeoSPARQL, but the integration is ontology-mediated and not released as a benchmark with tasks, splits, and baselines. KCityChatBot [38] pairs a CityGML KG with a multi-agent LLM pipeline but provides no reusable benchmark. On the systems side, 3DCityDB [62, 63] is the most

Table 1: *AuthentiCity* and related datasets and benchmarks. The comparison is on capability coverage; see the note below on suite size.

Resource	Spatial	Multi	Prov	Infeas	RL
Text-to-query benchmarks					
Spider [65]	✗	✗	✗	✗	✗
BIRD [36]	✗	✗	✗	✗	✗
NL2SQL-BUGs [39]	✗	✗	✗	(✓)	✗
Text2Cypher [50]	✗	✗	✗	✗	✗
CypherBench [21]	✗	✗	✗	✗	✗
Mind the Query [14]	✗	✗	✗	✗	✗
Urban data resources and benchmarks					
CityBench [20]	(✓)	✗	✗	✗	✗
UUKG [47]	(✓)	✗	✗	✗	✓
Ding et al. [16]	✓	✓	✗	✗	✗
AuthentiCity (+*pykci* [45])	✓	✓	✓	✓	✓

Spatial: predicates over 2D/3D geometries. Multi: multiple sources adding to the same entities. Prov: explicit provenance representation and provenance-aware evaluation. Infeas: deliberately unanswerable queries. RL: representation learning tasks. Parenthesized (✓) marks partial or narrower support. The table compares **capability coverage**, not suite size: *AuthentiCity* releases 1,394 executable, gold-verified questions, against 10,181 for Spider and more than 10,000 for CypherBench over far smaller, non-geometric graphs.

widely adopted CityGML platform, and Semantic Web representations have been studied extensively [12], but neither natively supports confidence-weighted cross-source correspondences or fact-level provenance, the capabilities central to our tasks (Section 4). *AuthentiCity* is complementary to these systems.

2.3 Multi-Source Integration and Matching

Fusing crowd-sourced and authoritative geodata requires entity resolution across polygon datasets. Optimal many-to-many polygon matching under the Jaccard measure is NP-hard [42], with scalable formulations building on tree-constrained bipartite matching [11, 43]. Rather than commit to a single set of hard matches, *AuthentiCity* retains the full weighted overlap graph as first-class confidence-annotated edges, letting queries and learning methods resolve correspondences as needed (Section 3); the released correspondences are also a resource for polygon-matching research, where ground-truth labels remain scarce.

2.4 Positioning

Table 1 summarizes the gap in existing datasets and benchmarks. Each row represents a strong benchmark within a well-established research area, yet, to the best of our knowledge, no existing benchmark combines these dimensions. *AuthentiCity* addresses this gap. To facilitate comparison with the closest prior work, we adopt evaluation metrics compatible with CypherBench and BIRD, particularly the execution accuracy.

3 The AuthentiCity Dataset

The *AuthentiCity* artifact. Five city-scale knowledge graphs: 180 GiB, 180M nodes, 220M edges, 1.2B properties, and 3.6M buildings.

Archive. DOI `10.5281/zenodo.21547211`

Code. https://github.com/hcu-cml/authenticity

Contents. Neo4j dump and backend-neutral node/edge export, loaders, question suite with gold queries and materialized answers, task splits, evaluation harness, and datasheet (Appendix A).

Licenses. Per source (Table 8); OpenStreetMap under ODbL [49]; our annotations and question suite under CC BY 4.0.

This section introduces *AuthentiCity*'s five datasets and its approach to multi-source integration and provenance management. An overview is provided in Figure 2 and Table 2.

3.1 Source Layers and the Provenance Spectrum

AuthentiCity is organized around an explicit *provenance spectrum* (Figure 2): every fact belongs to one of four trust classes recoverable directly from the graph structure rather than external documentation. Three rules enforce this. First, a derived value never overwrites an authoritative one; it is stored under a source-specific property (`osm_height` alongside the surveyed `measured_height`), so both are comparable in one query. Second, each external object is a separate node that retains its source identity, metadata, and geometry, linked to the authoritative anchor by an explicit edge rather than merged into it. Third, every fusion edge stores its match confidence, so consumers make their own trust decisions rather than inheriting a fixed resolution.

3.2 Authoritative Layer and Graph Construction

The authoritative layer is derived from open-government CityGML 2.0 LoD2 datasets, using Hamburg's state mapping release [41] with ALKIS cadastral semantics. *pykci* transforms CityGML into a compact Neo4j LPG [51], where semantically meaningful elements become nodes, syntactic wrappers are merged into edges, and coordinates are preserved verbatim. Each top-level feature is additionally registered in an R-tree spatial index [54], enabling semantic traversals and spatial predicates to be combined in a single query. The transformation is idempotent (i.e., re-ingest yields an identical graph) and dataset-independent. Mapping details and losslessness evaluation are reported in the companion system paper [45].

CityGML datasets use different coordinate reference systems (CRS), including national metric grids (Hamburg, Zurich, and Helsinki), geographic coordinates (Tokyo), and US survey feet (New York). Rather than enforcing a common CRS, each city is maintained in a local metric CRS to preserve accurate distance and area measurements. OSM data are reprojected into the corresponding city-specific CRS before fusion.

3.3 Identity Resolution and Integrity Check

Several source datasets violate the assumption that `gml:id` values are unique, including 100 duplicated identifiers in Zurich and Helsinki and more than 975,000 (33 % of input) ward-boundary identical buildings in Tokyo. We resolve these during ingestion without

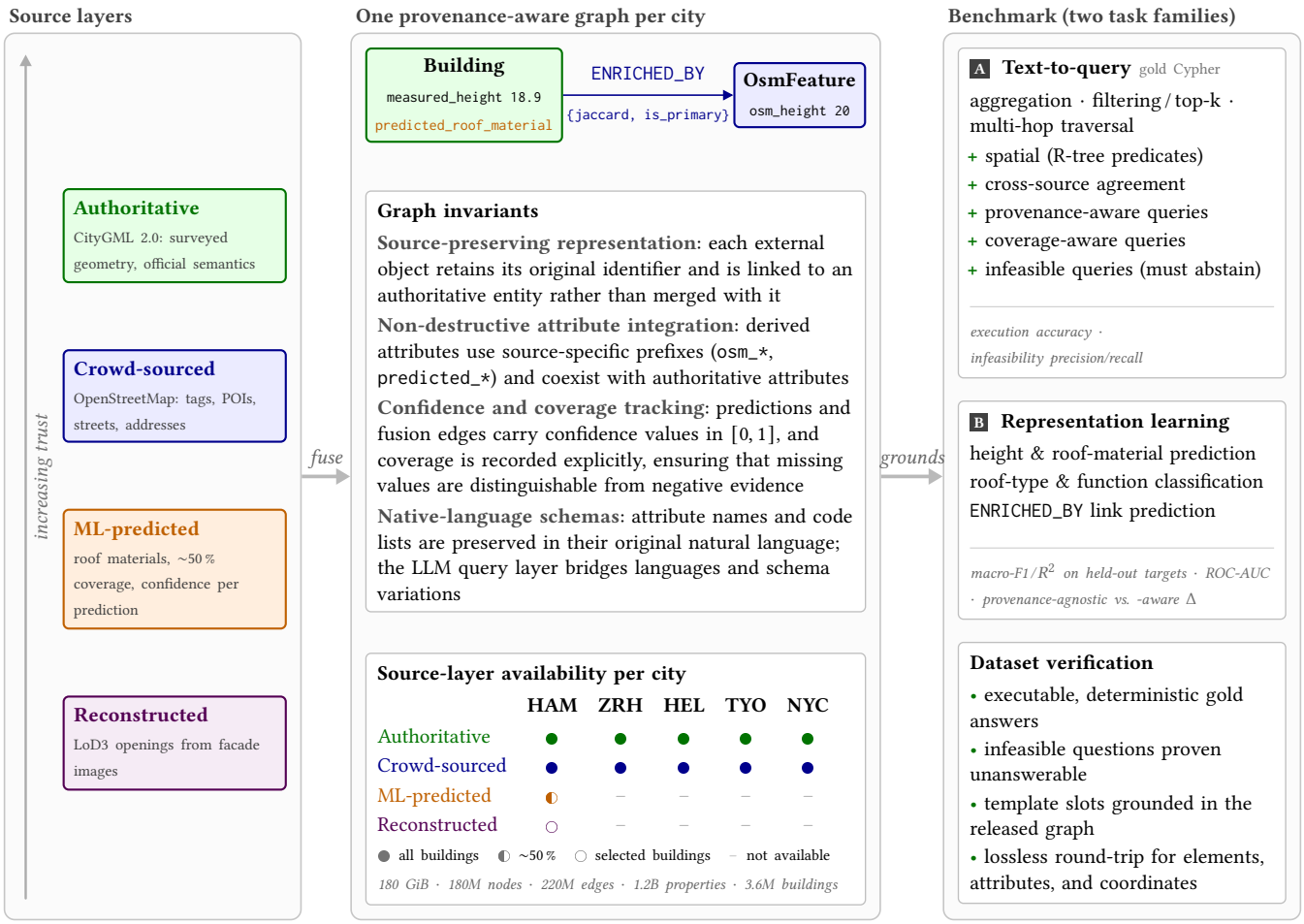


Figure 2: Overview of *AuthentiCity*. Four source layers with different trust levels (left) are integrated into a provenance-aware LPG for each city (center). The resulting graphs support two benchmark families with task-specific metrics and dataset-level validation (right). Categories marked (+) are absent from existing text-to-query benchmarks.

Table 2: Overview of *AuthentiCity*'s five city-scale datasets (full statistics in Section D, Tables 8 and 10). Hamburg is the Tier 2 deep-fusion instance; the remaining cities form the Tier 1 comparable core.

	Hamburg DE	Helsinki FI	Zurich CH	New York US	Tokyo JP	All 5 corpus
CityGML / LoD	2.0 / LoD2	2.0 / LoD2	2.0 / LoD2.3	1.0/2.0 / LoD1–2	2.0 / LoD1–3	–
Metric CRS	EPSG:25832	EPSG:3879	EPSG:2056	EPSG:32618†	EPSG:6677†	–
Buildings	388,267	2980	102,668	1,083,437	2,005,762	3,583,114
Nodes	17.44 M	0.41 M	41.51 M	45.80 M	74.41 M	179.57 M
Edges	26.30 M	0.62 M	45.55 M	54.21 M	91.95 M	218.62 M
Nodes/bldg. (med.)	30	54	253	31	21	–
Thematic keys	19	49	22	5	20	–
OSM-enr. (%)	84.9	81.1	91.2	98.7	57.8	74.1
Store (GiB)	14.8	0.3	30.2	32.4	99.2	176.9
Dump (GiB)	3.1	0.2	5.6	7.0	11.3	27.2

Nodes/bldg. is the median node count per building. *Keys* counts distinct thematic property keys. *OSM-enr.* is the share of buildings matched to at least one OSM feature. *Store*/*Dump* report Neo4j storage size and compressed release size. † Reprojected at ingest from a non-metric CRS.

modifying the source files: features sharing an identifier are compared using content and geometry hashes, with identical features merged and distinct features assigned unique identifiers, preventing both erroneous merges and artificial duplication. We further perform pre-ingestion validation to detect identifier, geometry, coordinate, and CRS anomalies, followed by post-ingestion checks of graph integrity, including resolved object counts, identifier uniqueness, provenance completeness, and spatial-index coverage.

3.4 Crowd-Sourced Layer: OSM Fusion

OSM provides an independent description of the city. Fusion focuses on footprint correspondence, where one building may map to multiple polygons in either source (1:1, 1:n, n:1, n:m). Candidate pairs A, B come from R-tree intersections and are retained when $r = \text{area}(A \cap B)/\min(\text{area}(A), \text{area}(B)) \geq \tau$. The overlap coefficient r serves as the acceptance criterion ($\tau = 0.3$). Each accepted edge stores additionally the Jaccard index $\text{area}(A \cap B)/\text{area}(A \cup B)$. Connected components of the bipartite overlap graph define correspondence classes, and component-level rules propagate source-prefixed OSM attributes to matched buildings (Section C, Figure 4). Geometry and points of interest remain on OSM nodes. Section C explains how the value of τ is selected. Unlike traditional conflation pipelines, we retain the full weighted overlap graph as `ENRICHED_BY` edges rather than enforcing a final matching, which is an NP-hard problem [42]. All OSM features are additionally registered in a second, dedicated R-tree layer, so crowd-sourced and authoritative geometry remain independently and jointly queryable.

3.5 ML-Predicted and Reconstructed Layer

The ML layer adds roof-material classes (concrete, metal, glass, roof tiles, tar paper) predicted from aerial orthophotos [4]. *AuthentiCity* preserves each building's full material distribution, storing every predicted class with its pixel coverage. Fusion is a direct ID join, with all values stored under source-prefixed keys. Predictions cover 194,799 of 388,267 buildings (50.2 %), limited by orthophoto availability.

The reconstructed layer adds LoD3 building models derived from facade imagery for 17 buildings in Hamburg's HafenCity district. The models contribute 416 windows and 83 doors as explicit `Opening` nodes, represented by 499 interior rings in wall geometries. Corrected facades are linked to their LoD2 buildings via `HAS_LOD3_FACADE` edges as separate nodes in a dedicated spatial layer, never replacing the measured geometry.

3.6 Cross-Source Agreement and Conflict

AuthentiCity's datasets are highly diverse (see Section D, Figure 5). Fusing these urban data sources creates both coverage and redundancy. We analyze geometric correspondences and dual-sourced attributes in Hamburg, Helsinki, Zurich, New York, and Tokyo.

Geometric disagreement. Most matched buildings align cleanly: 87.2 % of Hamburg's 293,895 correspondences are 1:1 matches (Table 11). In 1:n components, one OSM building covers 2.5 CityGML buildings on average; in n:1 components, one CityGML building corresponds to 3.0 OSM buildings. The 2363 n:m components motivate retaining the full overlap graph rather than forcing a single matching. Fragmentation also varies by city, from 0.4 % in New York to 9.8 % in Tokyo.

Table 3: Footprint correspondence between CityGML and OSM (share of components per city). Full counts in Table 11.

City	1:1	1:*n*	*n*:1	*n*:*m*	Fragmented
Hamburg	87.2	6.5	5.5	0.8	6.3
Helsinki	84.2	6.3	6.4	3.1	9.5
Zurich	90.8	5.8	2.4	0.9	3.3
New York	99.3	0.3	0.3	0.1	0.4
Tokyo	85.9	4.3	2.3	7.5	9.8

Fragmented comprises the *n*:1 and *n*:*m* cases, that is, CityGML buildings split across multiple OSM building footprints.

Table 4: OSM ingestion census.

City	Ingested	Coverage	Anchored	Standalone
Hamburg	1,188,139	95.6 %	356,707	831,432
Helsinki	55,930	89.7 %	4281	51,649
Zurich	1,860,401	95.8 %	102,794	1,757,607
New York	2,555,468	97.6 %	1,121,901	1,433,567
Tokyo	3,830,408	97.9 %	1,235,353	2,595,055

Ingested is how many OSM features enter the graph, not including non-network lines like barriers, building-annotation open ways, and man-made linear features. *Coverage*: ingested features as a share of all source OSM features in the city's bounding box. *Anchored* vs. *Standalone* split the ingested features by whether they attach to a CityGML feature or are kept as net-new nodes.

Attribute agreement and conflict. Height and storey counts largely agree when they are present (Table 12): storeys match exactly for 87.2 % of 144,364 buildings (98.6 % within ±1), and the median height difference over 3404 buildings is 0.27 m. Disagreements remain informative, including 219 buildings differing by more than 5 m. Cross-city behavior differs substantially: Zurich shows much lower height agreement (median $|\Delta h| = 2.3$ m), likely reflecting differing measurement conventions.

Roof material provides the richest comparison. Among 3641 buildings with both ML and OSM labels, agreement reaches 75.2 %. Agreement is highest for roof tiles (92.6 %) and lower for flat-roof materials. OSM also contributes 5059 roof-material values without predictions and 134 values outside the prediction taxonomy.

3.7 Multilingual Urban Knowledge Graphs

Helsinki, Hamburg, and especially Tokyo illustrate why natural-language interfaces matter for urban KGs. Tokyo's PLATEAU data uses Japanese attribute names and values, for example 地区計画 (district plan) and 市谷柳町地区, which *pykci* preserves verbatim as Unicode property keys. Traditionally, querying such data required familiarity with both the local schema and natural language. With LLM-grounded querying, users can ask questions in their own language, as shown in this work, making Tokyo's 100 GiB graph accessible to both local residents and international analysts.

4 Benchmark A: Natural-Language-to-Query

To show that the released graphs can be queried in natural language, and that doing so demands reasoning beyond what conventional text-to-SQL and text-to-Cypher benchmarks test, we define a natural-language-to-query benchmark. Given a natural-language question and the graph schema, a model must either produce an executable Cypher query whose result matches the gold answer or explicitly declare the question infeasible. We evaluate on LPGs by design. While a relational CityGML store can support aggregate, filter, and spatial queries, the benchmark's cross-source agreement, provenance- and confidence-filtered retrieval, and coverage-aware aggregation rely on LPG-native fusion structures absent from relational schemas.

4.1 Categories

We design 1394 query questions spanning 84 templates and nine categories (Table 14; full inventory in Section F), five of which are our contributions. *Spatial* questions, the largest category, cover window, proximity, nearest-neighbor, geometric multi-hop, 3D-structure, and density queries. *Cross-source* questions compare authoritative and crowd-sourced values and ask what the crowd-sourced layer reports where authoritative data is absent, a common case in New York, where 97.5 % of buildings have an OSM height but none an authoritative one. *Provenance-filtered* questions constrain answers by source or confidence, including traps ("using only authoritative data…") where reading a crowd-sourced value yields a plausible but incorrect answer. *Coverage-aware* questions require recognizing that aggregates over partially covered attributes are meaningful only with respect to the covered subset; the flagship trap asks for a city-wide roof-material count, where the gold answer couples count and coverage. *Infeasible* questions have no valid answer and test whether a model declines rather than fabricates a query, each paired with a guard query proving the information absent.

Two design choices increase difficulty. First, as feasibility depends on city-specific coverage, the same template can be coverage-aware in one city and infeasible in another (e.g., roof material is a coverage question in Hamburg but infeasible elsewhere). Second, ten templates are additionally posed in German and Japanese, including three Tokyo templates whose gold queries filter on Japanese property keys. A thematic subset targets city-specific schema content, including Helsinki's floor area, volume, and construction year; Zurich's data-vintage fields; and Tokyo's Japanese-keyed attributes (Section E).

4.2 Construction, Verification, and Metrics

The benchmark is generated from 84 schema-grounded templates, each paired with a hand-authored gold Cypher query whose slot values are read directly from the released graph. Feasible templates declare the graph tokens they require and instantiate only where those tokens exist; missing attributes are exercised through the infeasible subfamily. Every question passes a five-stage verification gate: gold queries execute, are deterministic, and are non-empty unless permitted; infeasible questions include a guard query; and each rebuild re-materializes all gold answers. This extends the methodology of [14]. Additionally, 282 questions (20.2 %) were manually reviewed, revealing a proximity-predicate defect that escaped automated checks and was fixed before release.

The benchmark is split by template into a public development set and a held-out test set with unpublished gold queries. The test set contains unseen templates and holds out 370 of 1394 questions (26.5 %). We report execution accuracy (result-set equivalence under canonicalization), decomposed into executed-and-correct (EX), executed-but-wrong, errored, and over-refusal. This distinction separates Cypher-generation failures from semantic failures. We also report infeasibility precision, recall, and F1, following prior NL-to-Cypher benchmarks [21, 36].

4.3 Setup and Results

We evaluate a local open-weight model (qwen2.5-coder:7b, Ollama `Q4_K_M`) and a commercial frontier model (Claude Sonnet 5, run as a closed-book Claude Code subagent). Both models receive byte-identical, schema-only context with no database or tool access. Outputs are cached and scored offline against the gold queries (examples in Section F.5). Even the frontier model leaves substantial headroom. Claude achieves 54–69 % EX across the five cities (Table 5). Nearly all remaining cases are *executed-but-wrong* rather than errored, indicating semantic rather than syntax failures. qwen reaches only 6–19 % EX and is dominated by *errored* queries (40–54 %), largely Cypher syntax errors in spatial tasks where it hallucinates PostGIS `ST_*` functions. This decomposition reveals qualitatively different failure modes that aggregate EX obscures. On infeasibility, qwen never refuses (recall 0), generating a query for every unanswerable question, whereas Claude declines appropriately (Inf-F1 72–95 on dev). Declines on New York and Tokyo stem primarily from spatial-query timeouts on the largest graphs (1–2M buildings) rather than semantic drift.

5 Benchmark B: Representation Learning

To demonstrate that the graph supports applications beyond text-to-query, including urban analytics such as energy efficiency [64] and urban planning [37], as well as provenance-aware tasks such as data fusion, quality assurance [5, 7], and cross-city transfer that underpin urban digital twins [1], we provide an embedding-based evaluation following the benchmark protocol of Dwivedi et al. [18]. It comprises three tasks under spatial-block and cross-city splits, namely *attribute imputation* of ML-predicted roof material (available for only ∼50% of buildings [4]) and authoritative building height (T1), *node classification* of administrative building-function and roof-type classes (T2), and *link prediction* of held-out `ENRICHED_BY` correspondence edges (T3). Each task is run under a *provenance-agnostic* protocol that hides source distinctions and a *provenance-aware* one that exposes source types and fusion-edge confidences, for example through Jaccard-weighted message passing, and the difference between them measures whether provenance helps. To our knowledge this is the first such evaluation on a real city-scale multi-source graph, although label coverage varies by task, with height spanning all five cities, matching per city, roof type in Hamburg and Helsinki, and building function only Hamburg (Section 5.2).

Table 5: Task family A results on all five city dev and held-out test splits (%). The four feasible outcomes sum to 100: EX (correct), Ex-wr (executed but incorrect), Err (execution failure), and O-ref (incorrect refusal). Inf-F1 measures infeasibility detection.

		Development set					Held-out test set				
Model	City	EX↑	Ex-wr	Err	O-ref↓	Inf-F1	EX↑	Ex-wr	Err	O-ref↓	Inf-F1
Claude	Hamburg	60.7	32.6	0.0	6.7	72.0	58.6	41.4	0.0	0.0	100.0
Sonnet 5	Helsinki	56.1	39.0	0.0	4.9	78.3	42.9	57.1	0.0	0.0	66.7
	Zurich	61.5	38.5	0.0	0.0	94.7	60.0	40.0	0.0	0.0	100.0
	New York	69.2	28.2	0.0	2.6	90.0	50.0	35.7	14.3	0.0	100.0
	Tokyo	53.7	37.3	7.5	1.5	90.0	41.7	37.5	20.8	0.0	100.0
qwen2.5-	Hamburg	19.1	36.0	44.9	0.0	n/a	20.7	48.3	31.0	0.0	n/a
coder:7b	Helsinki	14.6	45.1	40.2	0.0	n/a	10.7	53.6	35.7	0.0	n/a
	Zurich	9.2	41.5	49.2	0.0	n/a	15.0	40.0	45.0	0.0	n/a
	New York	10.3	35.9	53.8	0.0	n/a	14.3	28.6	57.1	0.0	n/a
	Tokyo	6.0	43.3	50.7	0.0	n/a	12.5	33.3	54.2	0.0	n/a

Inf-F1 is the F1 score for infeasibility detection, where over-refusals count as false positives. qwen2.5-coder never refuses, so Inf-F1 is undefined (n/a) and recall is 0. For qwen2.5-coder, most Err cases are Cypher syntax errors in spatial queries caused by hallucinated PostGIS `ST_*` functions. For Claude, most Err cases on the NYC and Tokyo test splits are spatial-query timeouts on the largest graphs (Section 4.3). Baselines are evaluated on a stratified subsample of the released suite; per-split counts are in the repository.

5.1 Provenance-Agnostic vs. Provenance-Aware

The two protocols differ only in how much of the provenance structure the encoder may use, so that any performance gap is attributable to provenance awareness rather than to model capacity. The *provenance-agnostic* setting flattens the multi-source graph into canonical entities by merging source-specific evidence, removing edge-confidence values, and discarding source labels. The *provenance-aware* encoder instead retains (i) source-typed nodes and relations (CityGML, OSM, ML-derived), (ii) confidence scores as edge weights, and (iii) coverage and agreement node features (source presence, count, and best-match confidence). Specifically, a canonical entity c aggregates evidence from its neighbors $e \in \mathcal{N}(c)$ as

$$\mathbf{h}'_c \;=\; \sigma\Big(\mathbf{W}_{\text{self}}\,\mathbf{h}_c \;+\; \sum_{e\in\mathcal{N}(c)} \frac{w_{ce}}{\sum_{e'} w_{ce'}}\,\mathbf{W}_{s(e)}\,\mathbf{h}_e\Big), \tag{1}$$

where $w_{ce} \in [0,1]$ is the correspondence confidence (the Jaccard overlap stored on the CityGML–OSM edges, normalized per target node) and $\mathbf{W}_{s(e)}$ is a source-specific projection, instantiated as one relation-specific weight matrix per typed relation. The agnostic variant sets $w_{ce} = 1$ and $\mathbf{W}_{s(e)} = \mathbf{W}$, recovering mean-pooled GraphSAGE [25].

5.2 Experimental Setup

Models. At full scale we evaluate an attribute-only MLP, provenance-agnostic GraphSAGE [25] and GAT [56], source-typed R-GCN [52], HAN [59], and HGT [27], the confidence-weighted encoder of Eq. (1), self-supervised DGI [57], and non-learned spatial-distance and attribute-similarity baselines for T3. The shallow node2vec [23] and metapath2vec [17] baselines are transductive and cannot embed unseen entities, so we exclude them from the full-scale comparison.

Splits and protocol. The *spatial* split holds out entire spatial tiles, and because a single hold-out is high-variance at city scale we use *spatial K-fold cross-validation* (K=5) with out-of-fold predictions pooled so every entity is tested once. The *cross-city* split is leave-one-city-out. Shallow methods are trained unsupervised, frozen, and probed, whereas GNNs are trained end-to-end. We report mean±std over 3 seeds for the survey (Table 6) and over 10 seeds for the cross-city result. For the focused provenance-agnostic versus provenance-aware paired comparisons and the confidence-weighting ablation, we repeat over 10 seeds and pair runs by seed, reporting a paired t-test [53], the Wilcoxon signed-rank test [61], Cohen's d_z [33], and a 95% confidence interval on the per-seed difference, and we treat a comparison as a null when the interval contains zero or the effect size is negligible. Provenance-agnostic and aware variants share backbone, depth, and budget (hidden dimension 64, 2 message-passing layers, dropout 0.3, Adam at learning rate 0.01, 150 epochs), and are trained in PyTorch Geometric on an Nvidia RTX PRO 6000 (96 GB). To prevent leakage, each predicted attribute is removed from the inputs, and a dedicated ablation additionally removes its correlated counterpart, such as storey count when predicting height. Data are streamed from Neo4j, with the loader handling missing node types, features, and labels so the same code runs unchanged across cities.

5.3 Experimental Results

A single-city subset (Hamburg, $n = 81$ buildings) first validated the pipeline and previewed both headline findings, that topology adds signal beyond attributes (with the collinear storey feature withheld, the attribute probe falls to $R^2 = -0.86$ while GraphSAGE recovers 0.52) and that provenance-awareness gives no measurable single-city benefit, including no gain on the seven buildings with conflicting OSM/CityGML roof evidence.

Full run (Hamburg, n=388k buildings). Within a single city, the aware encoder's advantage over the agnostic baseline is statistically consistent but negligible in magnitude. Over 10 seeds it improves T1 height (0.729 → 0.733), T2 roof type (0.556 → 0.558), and T2 building function (0.464 → 0.474), each significant under a paired

t-test ($p < 0.02$) yet at most 0.010 in absolute terms (cf. Table 6, 3-seed run). R-GCN, source-typed but not confidence-weighted, stays within about 0.02 of the full encoder on all three tasks, which points to source typing rather than confidence weighting as the origin of even this small effect. We therefore state a falsifiable hypothesis: source typing and confidence weighting are separable, and confidence weighting is redundant with source typing in-distribution but not under distribution shift. We test it with an ablation that removes only the confidence weights, leaving source typing, coverage features, and architecture unchanged. In-distribution, removing confidence weighting changes every task by at most 0.001, with no significant effect on height ($p = 0.21$) or building function ($p = 0.59$) and only a negligible 0.0006 on roof type ($p = 0.02$), confirming that confidence weighting is redundant with source typing in-distribution. Whether it also contributes under distribution shift, where the aware encoder's cross-city gains appear (Table 7), requires the same ablation under leave-one-city-out and remains future work.

In the T3 matching embedding space (Fig. 6), the aware encoder separates OSM-matched buildings into a distinct region where the agnostic encoder does not, and the ML- and OSM-coverage gaps co-locate, indicating that the two missingness patterns are correlated rather than independent, a restructuring the supervised metrics alone do not reveal. Under leave-one-city-out on T1 height the aware encoder improves on four of the five held-out cities (Table 7), with Hamburg rising from 0.270 to 0.548 and its standard deviation dropping from 0.170 to 0.071, plus gains on Zurich, New York City, and Tokyo, and Helsinki tied within noise. Roof-type transfer, possible only on the Hamburg–Helsinki pair, shows no meaningful gap between the two encoders (Table 18). For T3 matching, every learned encoder falls far short of non-learned baselines, for a structural reason. `ENRICHED_BY` correspondences are defined by footprint overlap (median Jaccard 0.842), so on the default nearest-neighbor negatives a spatial-distance rule reaches ROC-AUC 0.95–0.996 while the encoders reach at most 0.57. On the ambiguous n:m subset, where a building overlaps several OSM candidates and distance is uninformative, the distance rule falls to 0.75 AUC, a trivial attribute rule (building height versus OSM levels) still reaches 0.94, and every learned encoder collapses to chance (0.47–0.49 AUC on 18,234 Hamburg cases). T3 is therefore not a coordinate lookup, but it exposes a concrete gap, namely that current encoders exploit geometry and ignore the cross-source attribute signal that actually disambiguates correspondences. We provide the distance and attribute heuristics as reference baselines and pose attribute-aware n:m matching as an open challenge.

Cross-source completion and auxiliary targets. Unlike the three survey tasks, roof-material imputation, which predicts the ∼50% of Hamburg buildings the external ML model did not label, gives the aware encoder its clearest single-city edge (macro-F1 0.283 → 0.316, concentrated in the `concrete` class). A circularity check confirms this is structural rather than leakage, as only 3.5% of labeled buildings carry a matched OSM roof-material tag.

6 Conclusion

We presented *AuthentiCity*, a multi-source, provenance-aware 3D city knowledge graph and a two-family benchmark on top of it.

Table 6: Full-scale (single-city) Hamburg results (388k buildings, spatial cross-validation, mean±std over 3 seeds). Height retains the collinear storey count.

Model	Roof type macro-F1 ↑	Bldg function macro-F1 ↑	Height R^2 ↑
MLP (attr-only)	0.504 ± .000	0.329 ± .000	0.666 ± .000
DGI (self-supervised)	0.516 ± .007	0.409 ± .002	0.683 ± .002
GraphSAGE (agnostic)	0.556 ± .000	0.462 ± .000	0.730 ± .002
GAT (agnostic)	0.531 ± .000	0.365 ± .001	0.676 ± .001
source-typed, not confidence-weighted			
HGT	0.541 ± .002	0.393 ± .007	0.714 ± .005
HAN	0.460 ± .003	0.208 ± .004	0.416 ± .026
R-GCN	0.558 ± .000	0.470 ± .003	0.713 ± .004
Conf.-weighted GNN (aware)	**0.559 ± .001**	**0.476 ± .001**	**0.732 ± .000**

Table 7: T1 cross-city height transfer, leave-one-city-out (R^2, mean±std over 10 seeds, best cross-city result per row in bold), where *own-city* is the single-city reference. The *probe* is a no-graph baseline pooled over the training cities and is not comparable to the single-city attribute probe of Table 6.

Held-out	n	probe	agnostic	aware	own-city
Hamburg	388,267	−0.145	0.270 ± .170	**0.548 ± .071**	0.733
Helsinki	2,980	−0.000	**0.516 ± .016**	0.500 ± .029	0.461
NYC	1,083,437	**0.258**	0.167 ± .034	0.188 ± .024	0.407
Tokyo	2,005,762	−0.484	0.011 ± .130	**0.082 ± .063**	0.399
Zurich	102,668	−0.011	0.581 ± .035	**0.601 ± .037**	0.269

The dataset makes origin, confidence, and coverage of urban facts explicit and queryable across authoritative, crowd-sourced, ML-predicted, and reconstructed layers. The benchmark evaluates capabilities that are not captured by existing text-to-query or urban reasoning benchmarks, including spatial reasoning over indexed 3D geometry, cross-source comparison, coverage-aware aggregation, infeasibility detection, and provenance-aware representation learning. We release the complete artifact under open licenses, including the enriched LPGs, question suite, gold queries, data splits, loaders, evaluation harness, and datasheet, together with an archival DOI. We hope *AuthentiCity* will serve both as a rigorous testbed for evaluating query-generation models on realistic, heterogeneous urban data and as a foundation for developing provenance-aware embedding methods, an important gap highlighted by our baseline results. The implementation, documentation, and additional resources associated with this project are publicly available at https://github.com/hcu-cml/authenticity.

References

[1] Mahmoud Abdelrahman, Edgardo Macatulad, Binyu Lei, Matias Quintana, Clayton Miller, and Filip Biljecki. 2025. What is a Digital Twin anyway? Deriving the definition for the built environment from over 15,000 scientific publications. *Building and Environment* 274 (2025), 112748. doi:10.1016/j.buildenv.2025.112748

[2] Jethro Akroyd, Sebastian Mosbach, Amit Bhave, and Markus Kraft. 2021. Universal Digital Twin - A Dynamic Knowledge Graph. *Data-Centric Engineering* 2 (2021), e14. doi:10.1017/dce.2021.10

[3] Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV). 2025. *Amtliches 3D-Gebäudemodell in der Ausprägung Level of Detail 2 (LoD2-DE).* https://www.adv-online.de/AdV-Produkte/Weitere-

Produkte/3D-Gebaeudemodelle-LoD/

[4] Lukas Arzoumanidis, Son H. Nguyen, Lara Johannsen, Filip Rothaut, Weilian Li, and Youness Dehbi. 2025. Object Detection for the Enrichment of Semantic 3D City Models with Roofing Materials. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* X-4/W6-2025 (2025), 9–16. doi:10.5194/isprs-annals-X-4-W6-2025-9-2025

[5] Filip Biljecki, Lawrence Zheng Xiong Chew, Nikola Milojevic-Dupont, and Felix Creutzig. 2021. Open government geospatial data on buildings for planning sustainable and resilient cities. *arXiv preprint arXiv:2107.04023* (2021). doi:10.48550/arXiv.2107.04023

[6] Filip Biljecki, Yoong Shin Chow, and Kay Lee. 2023. Quality of Crowdsourced Geospatial Building Information: A global Assessment of OpenStreetMap Attributes. *Building and Environment* 237 (2023), 110295. doi:10.1016/j.buildenv.2023.110295

[7] Filip Biljecki, Hugo Ledoux, Xin Du, Jantien Stoter, Kean Huat Soon, and Victor Khoo. 2016. The most common geometric and semantic errors in CityGML datasets. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. IV-2/W1. International Society for Photogrammetry and Remote Sensing (ISPRS), 13–22. doi:10.5194/isprs-annals-IV-2-W1-13-2016

[8] Bundesamt für Landestopografie swisstopo. 2022. *Nutzungsbedingungen für kostenlose Geodaten und Geodienste (OGD) von swisstopo.* https://www.swisstopo.admin.ch/de/nutzungsbedingungen-kostenlose-geodaten-und-geodienste

[9] Bundesamt für Landestopografie swisstopo. 2026. *swissBUILDINGS3D 3.0 Beta.* https://www.swisstopo.admin.ch/de/landschaftmodell-swissbuildings3d-3-0-beta

[10] Bundesrepublik Deutschland. 2026. *Datenlizenz Deutschland – Namensnennung – Version 2.0.* https://www.govdata.de/dl-de/by-2-0

[11] Stefan Canzar, Khaled Elbassioni, Gunnar W. Klau, and Julián Mestre. 2011. On Tree-Constrained Matchings and Generalizations. In *Automata, Languages and Programming*, Luca Aceto, Monika Henzinger, and Jiří Sgall (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 98–109.

[12] Arkadiusz Chadzynski, Nenad Krdzavac, Feroz Farazi, Mei Qi Lim, Shiying Li, Ayda Grisiute, Pieter Herthogs, Aurel von Richthofen, Stephen Cairns, and Markus Kraft. 2021. Semantic 3D City Database - An Enabler for a Dynamic Geospatial Knowledge Graph. *Energy and AI* 6 (2021), 100106. doi:10.1016/j.egyai.2021.100106

[13] Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei Lan, Sheng Zhang, Jiarong Jiang, Joseph Lilien, Steve Ash, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, and Bing Xiang. 2023. Dr.Spider: A Diagnostic Evaluation Benchmark towards Text-to-SQL Robustness. arXiv:2301.08881 [cs.CL] https://arxiv.org/abs/2301.08881

[14] Vashu Chauhan, Shobhit Raj, Shashank Mujumdar, Avirup Saha, and Anannay Jain. 2025. Mind the Query: A Benchmark Dataset towards Text2Cypher Task. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (Eds.). Association for Computational Linguistics, Suzhou (China), 1890–1905. doi:10.18653/v1/2025.emnlp-industry.133

[15] Creative Commons. 2026. *Creative Commons Attribution 4.0 International License.* https://creativecommons.org/licenses/by/4.0

[16] Linfang Ding, Guohui Xiao, Albulen Pano, Mattia Fumagalli, Dongsheng Chen, Yu Feng, Diego Calvanese, Hongchao Fan, and Liqiu Meng. 2025. Integrating 3D City Data through Knowledge Graphs. *Geo-spatial Information Science* 28, 2 (2025), 780–799. arXiv:https://doi.org/10.1080/10095020.2024.2337360 doi:10.1080/10095020.2024.2337360

[17] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 135–144. doi:10.1145/3097983.3098036

[18] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2023. Benchmarking Graph Neural Networks. *Journal of Machine Learning Research* 24, 43 (2023), 1–48. http://jmlr.org/papers/v24/22-0567.html

[19] Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 591–602. doi:10.1145/3711896.3736878

[20] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. CityBench: Evaluating the Capabilities of Large Language Models for Urban Tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 5413–5424. doi:10.1145/3711896.3737375

[21] Yanlin Feng, Simone Papicchio, and Sajjadur Rahman. 2025. CypherBench: Towards Precise Retrieval over Full-scale Modern Knowledge Graphs in the LLM Era. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 8934–8958. doi:10.18653/v1/2025.acl-long.438

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. doi:10.1145/3458723

[23] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 855–864. doi:10.1145/2939672.2939754

[24] Gerhard Gröger, Thomas H. Kolbe, Claus Nagel, and Karl-Heinz Häfele. 2012. *OGC City Geography Markup Language (CityGML) Encoding Standard.* Open Geospatial Consortium (OGC). https://portal.ogc.org/files/?artifact_id=47842 OGC 12-019, Version 2.0.0, International Standard.

[25] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 1024–1034.

[26] Helsingin kaupunginkanslia. 2022. *3D Models of Helsinki.* https://hri.fi/data/en_GB/dataset/helsingin-3d-kaupunkimalli

[27] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proceedings of The Web Conference 2020 (WWW)*. 2704–2710. doi:10.1145/3366423.3380027

[28] デジタル庁. 2024. 公共データ利用規約 *(第1.0版)*. https://www.digital.go.jp/resources/open_data/public_data_license_v1.0

[29] 国土交通省都市局. 2025. *3D都市モデル（Project PLATEAU）ポータルサイト*. https://www.geospatial.jp/ckan/dataset/plateau

[30] Elmehdi Kanna, Jannik Matijevic, Lukas Arzoumanidis, Huynh Duc An Son Nguyen, and Youness Dehbi. 2026. Semantic Enrichment of 3D City Models via Roof Material Classification for Urban Greening and Heat Island Mitigation. *SSRN Electronic Journal* (2026). doi:10.2139/ssrn.6127041

[31] Bernd Ketzler, Vasilis Naserentin, Fabio Latino, Christopher Zangelidis, Liane Thuvander, and Anders Logg. 2020. Digital Twins for Cities: A State of the Art Review. *Built Environment (1978-)* 46, 4 (2020), 547–573. http://www.jstor.org/stable/45299343

[32] Thomas H. Kolbe, Tatjana Kutzner, Carl Steven Smyth, Claus Nagel, Carsten Roensdorf, and Charles Heazel. 2021. *OGC City Geography Markup Language (CityGML) Part 1: Conceptual Model Standard.* Open Geospatial Consortium (OGC). https://www.opengis.net/doc/IS/CityGML-1/3.0 20-010, Version 3.0.0, International Standard.

[33] Daniel Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4 (2013), 863. doi:10.3389/fpsyg.2013.00863

[34] Binyu Lei, Patrick Janssen, Jantien Stoter, and Filip Biljecki. 2023. *Challenges of Urban Digital Twins: A Systematic Review and a Delphi Expert Survey.* Vol. 147. Elsevier BV, 104716. doi:10.1016/j.autcon.2022.104716

[35] Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin SU, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2025. Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows. In *International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (Eds.), Vol. 2025. 28691–28735. https://proceedings.iclr.cc/paper_files/paper/2025/file/46c10f6c8ea5aa6f267bcdabcb123f97-Paper-Conference.pdf

[36] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can LLM Already Serve as a Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 42330–42357. https://proceedings.neurips.cc/paper_files/paper/2023/file/83fc8fab1710363050bbd1d4b8cc0021-Paper-Datasets_and_Benchmarks.pdf

[37] Pengyuan Liu and Filip Biljecki. 2022. A review of spatially-explicit GeoAI applications in Urban Geography. *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), 102936. doi:10.1016/j.jag.2022.102936

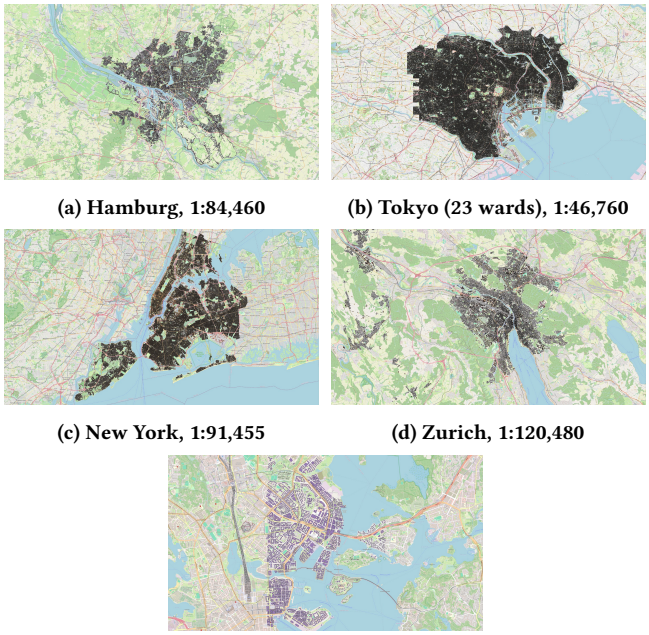
[38] S. Liu and C. Wang. 2025. KCitychatBot: A Knowledge Graph Based Chatbot System for Large-scale CityGML Dataset. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-4/W15-2025 (2025), 99–105. doi:10.5194/isprs-archives-XLVIII-4-W15-2025-99-2025

[39] Xinyu Liu, Shuyu Shen, Boyan Li, Nan Tang, and Yuyu Luo. 2025. NL2SQL-BUGs: A Benchmark for Detecting Semantic Errors in NL2SQL Translation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 5662–5673. doi:10.1145/3711896.3737427

[40] Yu Liu, Jingtao Ding, Yanjie Fu, and Yong Li. 2023. UrbanKG: An Urban Knowledge Graph System. *ACM Trans. Intell. Syst. Technol.* 14, 4, Article 60 (May 2023), 25 pages. doi:10.1145/3588577

[41] Metaver Metadatenverbund, Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg. 2025. *3D-Gebäudemodell LoD2-DE Hamburg.* https://metaver.de/trefferanzeige?cmd=doShowDocument&docuuid=2C1F2EEC-

CF9F-4D8B-ACAC-79D8C1334D5E
[42] Alexander Naumann, Annika Bonerath, and Jan-Henrik Haunert. 2024. Many-To-Many Polygon Matching à La Jaccard. In *32nd Annual European Symposium on Algorithms (ESA 2024) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 308)*, Timothy Chan, Johannes Fischer, John Iacono, and Grzegorz Herman (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 90:1–90:15. doi:10.4230/LIPIcs.ESA.2024.90
[43] Alexander Naumann, Annika Bonerath, and Jan-Henrik Haunert. 2025. Scalable Many-to-many Building Footprint Matching. *Information Fusion* 124 (2025), 103360. doi:10.1016/j.inffus.2025.103360
[44] New York City Office of Technology and Innovation (OTI). 2016. *3-D Building Model.* https://github.com/CityOfNewYork/nyc-geo-metadata/blob/main/Metadata/Metadata_3DBuildingModel.md
[45] Huynh Duc An Son Nguyen, Lukas Arzoumanidis, and Youness Dehbi. 2026. pykci: A Compact Urban Knowledge Graph for Semantic and Spatial Queries using LLMs. arXiv:2607.01605 [cs.DB] https://arxiv.org/abs/2607.01605
[46] Yansong Ning and Hao Liu. 2024. UrbanKGent: A Unified Large Language Model Agent Framework for Urban Knowledge Graph Construction. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 123127–123154. doi:10.52202/079017-3913
[47] Yansong Ning, Hao Liu, Hao Wang, Zhenyu Zeng, and Hui Xiong. 2023. UUKG: Unified Urban Knowledge Graph Dataset for Urban Spatiotemporal Prediction. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 62442–62456. https://proceedings.neurips.cc/paper_files/paper/2023/file/c4a30a4dd840cfeff30ba4d2661ff097-Paper-Datasets_and_Benchmarks.pdf
[48] NYC Office of Technology and Innovation (OTI). 2026. *NYC Open Data - Overview.* https://opendata.cityofnewyork.us/overview/
[49] Open Knowledge Foundation. 2026. *Open Data Commons Open Database License (ODbL).* https://opendatacommons.org/licenses/odbl/
[50] Makbule Gulcin Ozsoy, Leila Messallem, Jon Besga, and Gianandrea Minneci. 2025. Text2Cypher: Bridging Natural Language and Graph Databases. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, Genet Asefa Gesese, Harald Sack, Heiko Paulheim, Albert Merono-Penuela, and Lihu Chen (Eds.). International Committee on Computational Linguistics, Abu Dhabi, UAE, 100–108. https://aclanthology.org/2025.genaik-1.11/
[51] Ian Robinson, Jim Webber, and Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data.* " O'Reilly Media, Inc.".
[52] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web – 15th International Conference (ESWC) (Lecture Notes in Computer Science, Vol. 10843)*. Springer, 593–607. doi:10.1007/978-3-319-93417-4_38
[53] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25. doi:10.2307/2331554
[54] Craig Taverner and Andreas Berger. 2025. Neo4j Spatial. https://github.com/neo4j-contrib/spatial Accessed: June 10, 2026.
[55] United Nations Department of Economic and Social Affairs (UN DESA). 2019. *World Urbanization Prospects: The 2018 Revision.* United Nations. https://www.un-ilibrary.org/content/books/9789210043144
[56] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR).*
[57] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep Graph Infomax. In *International Conference on Learning Representations (ICLR).*
[58] Mohammad Saif Wajid, Hugo Terashima-Marin, Peyman Najafirad, Santiago Enrique Conant Pablos, and Mohd Anas Wajid. 2024. DTwin-TEC: An AI-based TEC District Digital Twin and Emulating Security Events by Leveraging Knowledge Graph. *Journal of Open Innovation: Technology, Market, and Complexity* 10, 2 (2024), 100297. doi:10.1016/j.joitmc.2024.100297
[59] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference (WWW)*. 2022–2032. doi:10.1145/3308558.3313562
[60] Zhu Wang, Fengxia Han, and Shengjie Zhao. 2024. A Survey on Knowledge Graph Related Research in Smart City Domain. *ACM Trans. Knowl. Discov. Data* 18, 9, Article 223 (Nov. 2024), 31 pages. doi:10.1145/3672615
[61] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. doi:10.2307/3001968
[62] Z. Yao, C. Nagel, M. Kendir, B. Willenborg, and T. H. Kolbe. 2025. The New 3D City Database 5.0 - Advancing 3D City Data Management based on CityGML 3.0. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* X-4/W6-2025 (2025), 241–248. doi:10.5194/isprs-annals-X-4-W6-2025-241-2025
[63] Zhihang Yao, Claus Nagel, Felix Kunde, György Hudra, Philipp Willkomm, Andreas Donaubauer, Thomas Adolphi, and Thomas H. Kolbe. 2018. *3DCityDB - A 3D Geodatabase Solution for the Management, Analysis, and Visualization of Semantic 3D City Models based on CityGML.* Vol. 3. Springer Science and Business Media LLC, 1–26. doi:10.1186/s40965-018-0046-7
[64] Winston Yap, Abraham Noah Wu, Clayton Miller, and Filip Biljecki. 2025. Revealing building operating carbon dynamics for multiple cities. *Nature Sustainability* 8, 10 (2025), 1199–1210. doi:10.1038/s41893-025-01615-8
[65] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3911–3921. doi:10.18653/v1/D18-1425



(a) Hamburg, 1:84,460 **(b) Tokyo (23 wards), 1:46,760**

(c) New York, 1:91,455 **(d) Zurich, 1:120,480**

(e) Helsinki (Kalasatama Digital Twins), 1:8734

Figure 3: Spatial extent of the five *AuthentiCity* cities. Basemap: © OpenStreetMap contributors.

A Datasheet

We document *AuthentiCity* following the datasheets framework of Gebru et al. [22]. Per-city figures referenced below are given in Tables 8 and 10. The cities' spatial extents are shown in Figure 3.

A.1 Motivation

For what purpose was the dataset created? To provide a multi-source, provenance-aware 3D city knowledge graph and a benchmark that evaluates reasoning over source origin, confidence, coverage, and cross-source agreement, which existing text-to-query and urban benchmarks do not exercise (Section 2).

Who created it? *AuthentiCity* was created by the authors of this paper as part of an academic research effort on multi-source urban data integration.

A.2 Composition

What do the instances represent? Nodes are buildings and other city objects (building parts, boundary surfaces, geometry primitives, and, per city, bridges, roads, vegetation, water bodies, and city furniture), together with fused OpenStreetMap features and, for Hamburg, ML-predicted roof materials and reconstructed LoD3 facades.

How many instances are there? Five city graphs totaling roughly 180 GiB, 180M nodes, 220M edges, 1.2B properties, and 3.6M buildings, ranging from 2980 buildings (Helsinki) to 2,005,762 (Tokyo).

Are relationships between instances made explicit? Yes, relationships are the dataset. Typed, directed edges connect each building to its parts, boundary surfaces, and geometry primitives, and to its district and city, while cross-source fusion is expressed as explicit `ENRICHED_BY`/`HAS_POI` edges carrying overlap and confidence attributes rather than as silently merged scalars (Section 3). Every edge records its source, so provenance is queryable at the relationship level, not just the node level.

Is any information missing, and does the dataset contain all instances or a sample? It contains all buildings of each released source extent; coverage of derived layers is deliberately partial and flagged (ML roof materials cover 50.2 % of Hamburg buildings; LoD3 covers 17 Hamburg buildings), so absence is never a negative label.

Does the dataset contain confidential or personal data? No. Instances are buildings, not people. All sources are already public. Address-level strings are retained only where OpenStreetMap itself publishes them. The remaining datasheet questions concerning human data subjects (consent, ethical review, data retention, offensive content) are therefore not applicable.

Are there errors, noise, or redundancies? Yes, and they are documented and preserved rather than silently altered. Some source datasets ship non-unique `gml:ids`, resolved at ingest by content and geometry hashing. Tokyo's PLATEAU uses a ±9999 height sentinel, kept verbatim but excluded from all statistics. The Helsinki source export carries upstream-corrupted Finnish diacritics (Section B).

A.3 Collection Process

How was the data acquired? Authoritative CityGML was downloaded from the open-government portals in Table 8. OpenStreetMap was obtained as regional Geofabrik extracts. The roof-material predictions are produced by our own imagery-based model [4], and the LoD3 facades are reconstructed from our own facade imagery.

Over what timeframe? The source vintages are listed per city in Table 8. The graph was constructed and fused in 2026.

A.4 Preprocessing, Cleaning, and Labeling

Was any preprocessing done? CityGML is mapped to a labeled property graph with coordinates stored verbatim. Non-metric sources (New York, Tokyo) are reprojected to a per-city metric CRS at ingest (Table 8). OpenStreetMap footprints are matched to authoritative footprints by a confidence-weighted overlap graph and attached as `ENRICHED_BY` edges without overwriting authoritative values. Derived scalars are prefixed (`osm_*`, `predicted_*`).

Is the raw source available? Yes, through the sources given in Table 8. The original coordinates, CRS, and `gml:ids` are retained as provenance so the mapping is auditable.

Was the data validated? Every released instance passes automated integrity gates: a lossless round-trip census on the authoritative layer and an OpenStreetMap ingestion-completeness census on the fused layer (Section 3.4).

A.5 Uses

Has the dataset been used for any tasks already? Yes, the two benchmark families reported in this paper, natural-language-to-query translation (Task A) and graph representation learning (Task B), are evaluated on it, and their baseline results are released alongside the dataset (Sections 4 and 5). We are aware of no third-party uses at time of release.

What tasks is the dataset intended for? The two benchmark families (Sections 4 and 5): natural-language-to-query translation and graph representation learning, both emphasizing provenance-, coverage-, and cross-source-aware reasoning. Additionally, the dataset can serve as a foundation for future graph-based urban analyses and data fusion.

What uses should be avoided? The dataset describes the built environment, not individuals, and should not be repurposed to infer information about residents. Derived layers are marked as predicted or reconstructed precisely so consumers can exclude them where authoritative-only evidence is required.

A.6 Distribution

How is the dataset distributed and under what license? As a Neo4j dump (27.2 GiB) plus a backend-neutral node/edge-table export, with loaders, task splits, gold queries and materialized answers, the evaluation harness, and this datasheet (Section 3). Each source retains its own license (Hamburg dl-de/by-2-0 [10], Zurich swisstopo [8], Helsinki CC BY 4.0 [15], New York NYC OpenData [48], Tokyo PLATEAU Public Data License 1.0 [28]). OpenStreetMap is licensed under the Open Data Commons Open Database License (ODbL) [49] by the OpenStreetMap Foundation (OSMF). Our annotations, question suite, and documentation are released under CC BY 4.0.

Is there a DOI? Yes, the artifact is archived on Zenodo under DOI `10.5281/zenodo.21547211`.

A.7 Maintenance

Who maintains it and how is it versioned? The authors host the leaderboard and version the suite. Any change to a gold answer on a dataset rebuild bumps the suite version under a changelog (gate G5, Section F.4).

Will it be extended? Yes: further Tier 1 cities and a sensed layer are planned as inference data becomes available (Section B).

Contact. The corresponding author of this paper, Huynh Duc An Son Nguyen (son.nguyen@hcu-hamburg.de).

B Limitations

Geographic scope. Tier 2 exists for one city and Tier 1 spans five across three continents. Findings may not transfer to regions with different cadastral traditions or OSM community density.

Cross-city heterogeneity. National schemes differ in detail and vocabulary, so cross-city questions use a common attribute subset (Section 3). Per-city attributes remain queryable but outside the comparable core. The binding constraint for the ML-predicted layer is inference-imagery availability.

Upstream source defects. Source values are preserved verbatim rather than silently repaired: the Helsinki export carries upstream-corrupted Finnish diacritics (1490 Unicode replacement characters across 283 values), which we retain as-is while the integrity gates confirm faithful preservation.

Prediction bias. The roof-material predictions inherit their training-imagery biases [4] and are marked as predicted.

C CityGML-OSM Knowledge Graph

Figure 4 illustrates the structure and content of *AuthentiCity*'s LPGs, which extend authoritative CityGML graphs with complementary OSM information. The enrichment process is strictly additive: authoritative CityGML entities and attributes are preserved, while OSM-derived information is attached either as source-prefixed properties on CityGML feature nodes or as separate OSM feature nodes linked through explicit relationships. Building-level correspondences between the two sources are represented through enrichment edges that record spatial matching statistics, including overlap ratio and Jaccard similarity, and identify the primary match based on the largest overlap. To avoid introducing ambiguity, only attributes from the primary OSM match are propagated to the corresponding CityGML building, while information from secondary matches remains associated with its original OSM feature. Throughout the graph, provenance is retained at the source level, enabling users to distinguish authoritative CityGML information from OSM-derived enrichments and trace the origin of all integrated data.

We choose τ (Section 3.4) through a sensitivity analysis of the correspondence structure. Since no city-scale ground-truth CityGML–OSM correspondences exist, we sweep $\tau \in [0.1, 0.7]$ and evaluate the mean Jaccard of primary matches (precision proxy) and the number of enriched buildings (recall proxy). The mean primary-match Jaccard remains constant (0.775), indicating that τ affects only marginal edges. We therefore select $\tau = 0.3$, which provides near-maximal enrichment while preserving correspondence quality. Because each retained edge stores its Jaccard score, downstream methods can reweight or rethreshold matches without repeating the alignment.

D Per-City Dataset and Cross-Source Analysis

This section gives an overview comparison of all five datasets Figure 5, the full per-city statistics summarized in Section 3, dataset provenance and scale (Table 8), graph size and content (Table 10), footprint correspondence between CityGML and OSM (Table 11), and dual-sourced attribute agreement (Table 12).

E Source Thematic Property Inventory

Every property key that originates in each city's source CityGML (core attributes plus `gen:*` generic attributes), with the number of buildings carrying it, is inventoried in Table 13. These verbatim, multilingual source vocabularies (German ALKIS/AdV codes, Finnish, Swiss-German, a sparse English LoD1–2 schema, and Japanese PLATEAU keys) are the semantics a text-to-query model must bridge (Section 4) and the attributes available for representation-learning tasks (Section 5).

F Question Suite Design and Statistics

This section summarizes the question suite specification. The full document (all 84 template definitions with gold-query sketches, slot providers, sizing model, and freeze process) ships with the benchmark repository. Table 14's instance counts are the frozen, live-verified v1.0 numbers.

F.1 Template Inventory

The categories of Table 14 are divided into capability subfamilies. The five distinctive categories are deliberately the deepest: spatial alone accounts for roughly a third of the feasible instances, and the five together for approximately 70 %. Templates marked cross-city use only the common attribute subset of Section 3 and instantiate on all five cities. City-specific templates (ALKIS code lists, Japanese-keyed attributes, prediction and LoD3 layers) are tagged as such and instantiate only where their layers exist, a constraint the generator asserts at run time.

F.2 Difficulty Rubric

Each question is assigned an authored difficulty level (Table 15). After completing the baseline evaluation matrix, we additionally report the observed difficulty of each question, measured by its failure rate across models, and quantify the correlation between authored and observed difficulty. Any discrepancy between the two is treated as an empirical finding rather than a reason to revise the original difficulty labels. Gold answers exhibit a diverse set of result formats (Table 16), determined by the structure of the corresponding gold query outputs. Single-value results are the most common, including aggregates, counts, and scalar lookups. These are followed by multi-column or grouped tables, ranked top-k lists, and infeasible cases that require refusal. Notably, none of the benchmark templates produces a bare boolean answer.

F.3 Per-City Feasibility

Table 17 specifies which benchmark categories are instantiated for each city. Missing entries are intentional design choices. In particular, the coverage and LoD3 rows define the Tier 2 deep-fusion categories. Likewise, the absent-layer infeasible subfamily relies on the absence of prediction layers outside Hamburg, allowing the same question template to be feasible in one city and infeasible in another while keeping the question text unchanged.

F.4 Verification Gate

Every question passes five machine-checked validation gates before entering the benchmark suite. **G1** verifies that the gold query

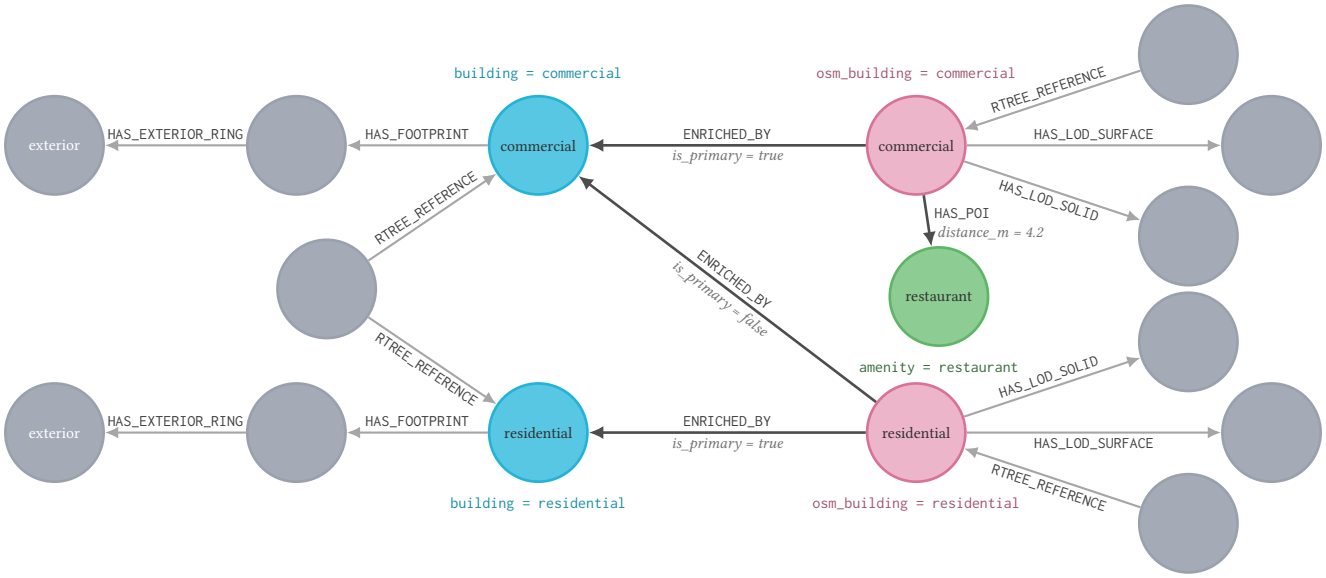


Figure 4: OSM enrichment: CityGML buildings (pink) are linked to OSM buildings (blue) by `ENRICHED_BY` edges. Buildings inherit source-prefixed `osm_*` attributes from their primary match, while points of interest (green) attach via `HAS_POI`.

Table 8: Dataset provenance and scale of the five *AuthentiCity* cities. Extent is the bounding box of all building footprints in the source CityGML (WGS84).

	Hamburg DE	Helsinki FI	Zurich CH	New York US	Tokyo JP	All 5 corpus
Provider (source)	LGV [41]	HRI [26]	swisstopo [9]	NYC OTI [44]	PLATEAU [29]	–
Data vintage	2025	2019	2019	2016	2025	2016–2025
License	dl-de/by-2-0 [10]	CC BY 4.0 [15]	swisstopo [8]	NYC OpenData [48]	PDL 1.0 [28]	–
CityGML / LoD	2.0 / LoD2	2.0 / LoD2	2.0 / LoD2.3	1.0/2.0 / LoD1–2	2.0 / LoD1–3	–
Metric CRS	EPSG:25832	EPSG:3879	EPSG:2056	EPSG:32618†	EPSG:6677†	–
Source files (tiles)	1[a]	1	78	20	2335	2435
Extent (km^2, bbox)	7117	14	3137	2159	1045	13,472
Source building features	388,267	2980	102,673	1,083,437	2,980,839	4,558,196
Distinct `gml:ids`	388,267	2919	102,628	1,083,437	2,005,294	3,582,545
Building nodes (graph)[b]	388,267	2980	102,668	1,083,437	2,005,762	3,583,114

[a] The Hamburg LGV publishes its LoD2 model as map tiles; the single file used here was produced by importing those tiles into 3DCityDB [62, 63] and re-exporting them as one merged CityGML, so ingest sees a single file whose bounding box is the union of the original tiles. [b] Nodes representing unique buildings, may differ from total number of input buildings, as in the case of Zurich and Tokyo. † Reprojected at ingest from a non-metric source CRS (Tokyo EPSG:6697 geographic degrees, New York EPSG:2263 US survey feet). Distinct `gml:id` counts differ from source feature counts where sources repeat IDs. Identity resolution at ingest splits distinct same-ID features onto synthesized IDs and deduplicates byte-identical replicas (Tokyo's ward-package tiling). Building node counts are thus lossless.

executes successfully without error. **G2** requires the canonicalized answer hash to be identical across three independent executions, ensuring result stability. **G3** checks that the result is non-empty unless the underlying template explicitly permits empty outputs. **G4**, applied to infeasible questions, executes a guard query that must demonstrate the absence of the requested information in the target city's graph by returning a count of zero. **G5** enforces benchmark versioning: whenever the dataset is rebuilt, all gold answers are re-materialized, and any change triggers a patch-version increment together with a changelog entry, preventing gold answers from silently drifting from the released graph. In addition to these automated checks, we perform human validation on a stratified sample comprising at least 20% of questions, sampled across categories, difficulty levels, and cities. Each sampled question is independently reviewed for gold-answer correctness and natural language quality. At benchmark freeze time, we document the rejection rates for each validation gate together with the human-review acceptance rates.

F.5 Example Questions with Gold Queries

One representative example is provided for each distinctive capability. Placeholders enclosed in angle brackets are instantiated at

Table 9: Multi-source enrichment of the five city graphs: crowd-source OSM, ML-predicted roof materials, and reconstructed LoD3 geometrical data.

	Hamburg DE	Helsinki FI	Zurich CH	New York US	Tokyo JP	All 5 corpus
Crowd-sourced layer (OSM)						
OSM features	1,188,139	55,930	1,860,401	2,555,468	3,830,408	9,490,346
Matched to a city object	356.7 k (30.0%)	4281 (7.7%)	102.8 k (5.5%)	1.12 M (43.9%)	1.24 M (32.3%)	2.82 M (29.7%)
Standalone net-new[a]	831.4 k (70.0%)	51.6 k (92.3%)	1.76 M (94.5%)	1.43 M (56.1%)	2.60 M (67.7%)	6.67 M (70.3%)
Buildings enriched (`ENRICHED_BY`)	331.3 k (85.3%)	2476 (83.1%)	94.0 k (91.5%)	1.07 M (98.8%)	1.18 M (58.9%)	2.68 M (74.8%)
`osm_*` property records copied	2.26 M	15.4 k	601.5 k	6.98 M	5.12 M	14.97 M
Buildings with ≥1 POI	12.0 k (3.1%)	657 (22.0%)	7569 (7.4%)	34.7 k (3.2%)	65.3 k (3.3%)	120.1 k (3.4%)
POI attachments (`HAS_POI`)	22,219	1797	13,587	52,449	91,487	181,539
ML-predicted layer (roof material; property records on Building nodes)						
Buildings with prediction	194.8 k (50.2%)	–	–	–	–	194.8 k (5.4%)
– thereof multi-material	570 (0.3%)	–	–	–	–	–
Material assignments (ranked)	195.4 k	–	–	–	–	195.4 k
Property records written	975.2 k	–	–	–	–	975.2 k
Reconstructed layer (LoD3)						
LoD3 facade surfaces	3404	–	–	–	–	3404
Anchor edges (`HAS_LOD3_FACADE`)	303	–	–	–	–	303

[a] Standalone nodes are net-new knowledge outside the cadastral layer, kept with an explicit `unmatched_reason`: features of a kind that has no CityGML counterpart to anchor to (roads, land use, water, street furniture), point features falling inside no building footprint (`no_containing_building`), and polygons overlapping no CityGML footprint (`no_overlap`), e.g. Hamburg: 752,624 / 56,715 / 22,093 of its 831,432 standalone features.

generation time using values retrieved directly from the released graph.

Spatial (S1 window). "Which is the tallest building inside the window ⟨wkt⟩?"

```
CALL spatial.intersects('features', '<wkt>') YIELD node
WITH node WHERE node:Building
  AND node.measured_height IS NOT NULL
RETURN node.id AS id, node.measured_height AS height_m
ORDER BY node.measured_height DESC, node.id ASC LIMIT 1
```

Cross-source (value disagreement). "Which buildings have an OSM height that differs from the surveyed height by more than ⟨*d*⟩ meters?"

```
MATCH (b:Building)-[r:ENRICHED_BY {is_primary: true}]
      ->(o:OsmFeature)
WHERE o.osm_height IS NOT NULL
  AND b.measured_height > -999
  AND abs(toFloat(o.osm_height) - b.measured_height) > <d>
RETURN b.id AS id, b.measured_height AS surveyed_m,
       toFloat(o.osm_height) AS osm_m
ORDER BY abs(toFloat(o.osm_height) - b.measured_height)
  DESC, id ASC
```

Provenance (source-restriction trap). "Using only authoritative cadastral data, what is the average building height?" The gold query reads the surveyed property only. A model that also averages the crowd-sourced `osm_height` produces a plausibly close but wrong number.

```
MATCH (b:Building)
WHERE b.measured_height > -999
RETURN round(avg(b.measured_height) * 100) / 100.0
       AS avg_height_m
```

Coverage-aware (city-wide count trap). "How many buildings in Hamburg have a ⟨material⟩ roof?" The gold answer couples the count with its coverage context. A bare count is scored as wrong.

```
MATCH (b:Building)
WITH count(b) AS total,
  count(CASE WHEN b.predictedroofmaterial IS NOT NULL
        THEN 1 END) AS covered,
  count(CASE WHEN b.predictedroofmaterial = '<material>'
        THEN 1 END) AS matching
RETURN matching, covered, total,
  round(1000.0 * covered / total) / 10.0 AS coverage_pct
```

Infeasible (absent attribute). "In which year was each building last renovated?" Gold behavior is an explicit refusal. The guard proves the absence on the target city:

```
MATCH (b:Building)
WHERE b.renovation_year IS NOT NULL
RETURN count(b) AS n  // gate G4: must return n = 0
```

Multilingual (Japanese-key schema bridging, Tokyo). "Which buildings belong to the district plan 'Ichigaya-Yanagichō'?", also posed in Japanese (`lang: ja`) as 「地区計画『市谷柳町地区』に属する建物はどれですか。」 The property key itself is Japanese, so the gold query must bridge the schema regardless of the question language:

```
MATCH (b:Building)
WHERE b.`地区計画` = '市谷柳町地区'
RETURN b.id AS id ORDER BY id
```

Table 10: Graph size and content (nodes and edges) of *AuthentiCity* KGs. Each cell denotes the absolute count and, in parentheses, its share of that city's total nodes or relationships. Predicted roof materials are stored as building properties.

	Hamburg DE	Helsinki FI	Zurich CH	New York US	Tokyo JP	All 5 corpus
Graph size						
Nodes	17,440,924	414,459	41,508,371	45,797,985	74,409,400	179,571,139
Relationships	26,299,391	619,964	45,548,405	54,207,134	91,949,234	218,624,128
Node property records	113.35 M	3.15 M	185.79 M	291.39 M	405.19 M	998.88 M
Relationship property records	25.60 M	513.1 k	39.33 M	46.18 M	73.39 M	185.01 M
Node content (count, % of the city's nodes)						
Semantic city objects	388.9 k (2.2%)	3050 (0.7%)	114.7 k (0.3%)	1.08 M (2.4%)	2.71 M (3.6%)	4.30 M (2.4%)
Buildings	388.3 k (2.2%)	2980 (0.7%)	102.7 k (0.2%)	1.08 M (2.4%)	2.01 M (2.7%)	3.58 M (2.0%)
Building parts	–	–	12.0 k (0.0%)	–	–	12.0 k (0.0%)
Other city objects[a]	–	31 (0.0%)	–	–	659.0 k (0.9%)	659.0 k (0.4%)
Interior & openings	499 (0.0%)	–	–	–	46.2 k (0.1%)	46.7 k (0.0%)
Classifier hubs	138 (0.0%)	39 (0.0%)	–	–	–	177 (0.0%)
Geometry & surfaces	15.84 M (90.8%)	354.6 k (85.6%)	39.50 M (95.2%)	42.10 M (91.9%)	67.77 M (91.1%)	165.57 M (92.2%)
Boundary surfaces	4.53 M (26.0%)	50.6 k (12.2%)	425.3 k (1.0%)	12.97 M (28.3%)	3.09 M (4.1%)	21.06 M (11.7%)
– roof	868.0 k (5.0%)	13.4 k (3.2%)	208.4 k (0.5%)	1.58 M (3.5%)	908.0 k (1.2%)	3.58 M (2.0%)
– wall	3.28 M (18.8%)	34.3 k (8.3%)	108.7 k (0.3%)	10.29 M (22.5%)	1.98 M (2.7%)	15.70 M (8.7%)
– ground	388.3 k (2.2%)	2923 (0.7%)	108.2 k (0.3%)	1.09 M (2.4%)	174.4 k (0.2%)	1.77 M (1.0%)
– closure/other	–	–	–	–	21.7 k (0.0%)	21.7 k (0.0%)
Polygons	4.95 M (28.4%)	141.1 k (34.0%)	19.25 M (46.4%)	14.34 M (31.3%)	29.56 M (39.7%)	68.25 M (38.0%)
Rings	4.96 M (28.4%)	142.1 k (34.3%)	19.26 M (46.4%)	14.35 M (31.3%)	29.60 M (39.8%)	68.31 M (38.0%)
Solids	388.3 k (2.2%)	5585 (1.3%)	108.7 k (0.3%)	–	2.18 M (2.9%)	2.68 M (1.5%)
Multi/composite surfaces	499 (0.0%)	31 (0.0%)	–	–	2.69 M (3.6%)	2.69 M (1.5%)
Line strings	619.4 k (3.6%)	15.2 k (3.7%)	455.5 k (1.1%)	446.0 k (1.0%)	643.1 k (0.9%)	2.18 M (1.2%)
Terrain intersections	388.2 k (2.2%)	–	–	–	–	388.2 k (0.2%)
Crowd-sourced (OSM features)	1.19 M (6.8%)	55.9 k (13.5%)	1.86 M (4.5%)	2.56 M (5.6%)	3.83 M (5.1%)	9.49 M (5.3%)
Reconstructed (LoD3 facades)	3404 (0.0%)	–	–	–	–	3404 (0.0%)
Containers & spatial index	20.4 k (0.1%)	893 (0.2%)	30.1 k (0.1%)	54.5 k (0.1%)	100.2 k (0.1%)	206.0 k (0.1%)
Dataset / City / District	3 (0.0%)	3 (0.0%)	157 (0.0%)	41 (0.0%)	3201 (0.0%)	3405 (0.0%)
R-tree index nodes	20.4 k (0.1%)	890 (0.2%)	29.9 k (0.1%)	54.4 k (0.1%)	97.0 k (0.1%)	202.6 k (0.1%)
Thematic (semantic) content[b]*: hub nodes vs. edges vs. source property records*						
Classifier hubs (function, roof type)	138	39	–	–	–	177
Classifier edges	776.5 k	3066	–	–	–	779.6 k
Distinct source thematic keys[c]	19	49	22	5	20	–
Thematic properties on buildings	6.97 M	104.5 k	1.94 M	5.42 M	31.51 M	45.94 M
Relationship content (count, % of the city's relationships)[d]						
Geometry composition	20.37 M (77.5%)	434.4 k (70.1%)	39.50 M (86.7%)	42.10 M (77.7%)	71.50 M (77.8%)	173.92 M (79.6%)
Semantic / thematic links	777.0 k (3.0%)	3066 (0.5%)	12.0 k (0.0%)	–	46.2 k (0.1%)	838.4 k (0.4%)
Topology, containment & index	4.75 M (18.1%)	177.7 k (28.7%)	5.92 M (13.0%)	10.97 M (20.2%)	18.92 M (20.6%)	40.74 M (18.6%)
Multi-source fusion	395.7 k (1.5%)	4808 (0.8%)	113.9 k (0.3%)	1.13 M (2.1%)	1.48 M (1.6%)	3.13 M (1.4%)
Per-building detail (median[e])						
Nodes / building	30	54	253	31	21	–
Geometry polygons / building	8	20	124	10	9	–
Source thematic attributes / building	18	41	20	5	16	–
Measured height[f] (m)	7.53	10.11	9.22	–	7.9	–

[a] Helsinki: 31 bridges. Tokyo (B-core module set): 605,293 roads, 38,231 city furniture, 10,883 vegetation objects, 968 bridges, 2895 water bodies, 735 plant cover, 1 city object group. [b] Excluding IDs, bounding boxes, spatial indices, and enriched properties (OSM, predicted roof materials, LoD3). [c] Property keys that originate in the input CityGML (core attributes and generic attributes), which may carry ML/DL-relevant semantics; Every key is inventoried per city in Table 13. [d] *Geometry composition*: `HAS_BOUNDARY`, `HAS_POLYGON`, `HAS_SURFACE_MEMBER`, `HAS_EXTERIOR_/INTERIOR_RING`, `HAS_LOD_SOLID/SURFACE/GEOMETRY`, `HAS_FOOTPRINT`, `HAS_LINE`, `HAS_TERRAIN_INTERSECTION`. *Semantic/thematic*: `HAS_FUNCTION`, `HAS_ROOF_TYPE`, `HAS_BUILDING_PART`, `HAS_OPENING`, installation/room/nesting edges. *Topology, containment, and index*: `PART_OF`, `LOCATED_IN`, `COVERS`, `RTREE_*` (spatial R-tree). *Multi-source fusion*: `ENRICHED_BY`, `HAS_POI`, `HAS_LOD3_FACADE`. [e] Per-building figures are medians (robust to a few hub `Building` nodes with heavy payloads). [f] Per-building `bldg:measuredHeight` attribute (height above ground per the national definition, not terrain elevation).

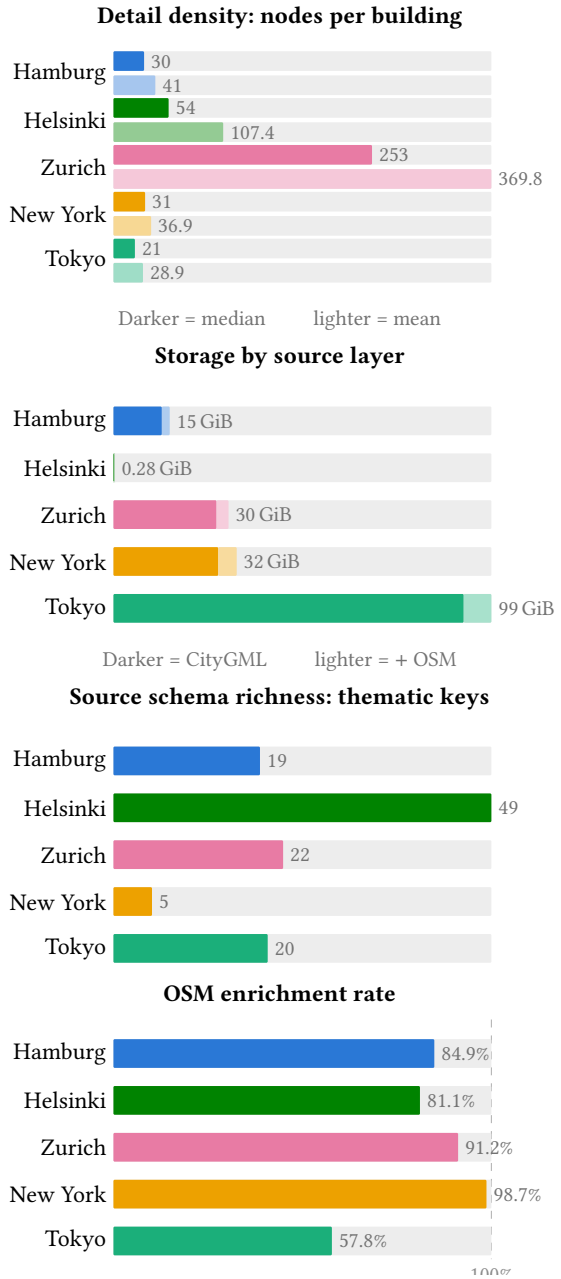


Figure 5: Comparison of the five *AuthentiCity* graphs. Zurich is the densest, Tokyo the largest, Helsinki the most attribute-rich, and New York the most OSM-enriched. Geometry accounts for 85–95 % of graph content in every city.

G LLM Setup for Text-to-Query

Both query-generation models are evaluated under identical, deliberately *closed-book* conditions. Each question prompt contains only the natural-language question, the project's Cypher and spatial-query rules, and a fixed per-city schema string consisting of the cached APOC-sampled graph schema together with the documented schema-gap patches (Section 4.3). No live database connection or external tool access is available. The schema string is byte-identical for both models within a given city, ensuring that neither

Table 11: Footprint correspondence between CityGML and OSM buildings.

Case	Components	Share	CityGML bldgs	OSM bldgs
Hamburg				
1:1	256,219	87.2 %	255,984	255,984
1:n	19,167	6.5 %	47,936	19,096
n:1	16,146	5.5 %	16,142	47,838
n:m	2363	0.8 %	9537	9151
Helsinki				
1:1	1692	84.2 %	1686	1686
1:n	127	6.3 %	285	125
n:1	129	6.4 %	127	281
n:m	62	3.1 %	318	302
Zurich				
1:1	75,541	90.8 %	75,430	75,430
1:n	4858	5.8 %	12,442	4853
n:1	2003	2.4 %	1999	4325
n:m	787	0.9 %	3812	3744
New York				
1:1	1,053,850	99.3 %	1,053,727	1,053,727
1:n	3046	0.3 %	8448	3015
n:1	3199	0.3 %	3178	6149
n:m	850	0.1 %	4277	4351
Tokyo				
1:1	823,560	85.9 %	821,655	821,655
1:n	41,392	4.3 %	90,648	40,561
n:1	22,288	2.3 %	22,273	46,975
n:m	71,562	7.5 %	224,653	224,226

Reading: in 1:n one OSM polygon covers several cadastral buildings; in n:1 several OSM polygons tile one cadastral building. The case mix is itself a per-city property spanning more than an order of magnitude: the fragmented share (n:1 plus n:m) is 0.4 % for New York (whose crowd mapping tracks the cadastre almost one-to-one), 3.3 % for Zurich, 6.3 % for Hamburg, 9.5 % for Helsinki, and 9.8 % in Tokyo, whose crowd mapping splits more cadastral buildings into several tagged parts.

Table 12: Dual-sourced building attributes when a CityGML, OSM or ML-predicted value describe the same property.

Property	City	Dual	Agreement or $\lvert\Delta\rvert$
Storeys	Hamburg	144,364	87.2 % exact; 98.6 % ±1
	Helsinki	48	70.8 % exact; 91.7 % ±1
	Tokyo	116,022	85.9 % exact; 94.3 % ±1
Height	Hamburg	3404	85.0 % ≤ 2 m; 6.4 % > 5 m
	Helsinki	76	61.8 % ≤ 2 m; 28.9 % > 5 m
	Zurich	770	46.4 % ≤ 2 m; 22.6 % > 5 m
	Tokyo	61,373	94.4 % ≤ 2 m; 2.5 % > 5 m
Roof mat.	Hamburg	3641	75.2 % over 5 classes

Table 13: Overview of every thematic property key (core and generic attributes) originating from input CityGML datasets with the number of buildings carrying it. Keys are verbatim source vocabulary in each portal's own language.

City	Source thematic property keys (buildings carrying the key)
Hamburg (19)	address_xml 388.3 k, creation_date 388.3 k, datenquellebodenhoehe 388.3 k, datenquelledachhoehe 388.3 k, datenquellelage 388.3 k, description 388.3 k, external_information_system 388.3 k, external_object_name 388.3 k, external_reference_xml 388.3 k, function_code 388.3 k, gemeindeschluessel 388.3 k, geometrietyp2dreferenz 388.3 k, grundrissaktualitaet 388.3 k, measured_height 388.3 k, measured_height_uom 388.3 k, roof_type_code 388.3 k, datenquellegeschossanzahl 375.7 k, storeys_above_ground 375.7 k, gml_name 1665
Helsinki (49)	creation_date 2980, description 2980, address_xml 2972, external_information_system 2972, external_object_name 2972, external_reference_xml 2972, groundlevel 2972, highestroof 2972, lowestroof 2972, measured_height 2972, measured_height_uom 2972, roof_type_code 2972, area_diff 2846, area_diff_filter 2846, file_candidate_gmlid 2846, integrating_person 2846, integration_date 2846, matching_mode 2846, overlap_db_to_file 2846, overlap_file_to_db 2846, overlap_filter 2846, code 2727, repaired 2727, uuid 2727, brec_buildingheightnn 2605, brec_roofnames 2605, gen_id 2176, rakennuksen_tila 2176, c_kayttark 1931, kerroksia 1931, kg_krakenn 1931, rakennustunnus__ratu 1931, rakennustunnus__vtj_prt 1931, tila_koodi 1931, kayttotarkoitus 1809, kerrosala 1809, valmistunut 1752, katuosoite 1750, tilavuus 1743, kokonaisala 1732, rakennusaine 1729, buildingheightnn 367, roofnames 367, suunnitelma_alue 245, kerrosala__m2 122, tila 122, tyyppi 122, function_code 94, storeys_above_ground 92
Zurich (22)	description 102.7 k, egid 99.3 k, dach_max 96.7 k, dach_min 96.7 k, datum_aenderung 96.7 k, datum_erstellung 96.7 k, erstellung_jahr 96.7 k, erstellung_monat 96.7 k, gen_predicate 96.7 k, grund_aenderung 96.7 k, herkunft 96.7 k, herkunft_jahr 96.7 k, herkunft_monat 96.7 k, objektart 96.7 k, original_herkunft 96.7 k, revision_jahr 96.7 k, revision_monat 96.7 k, gelaendepunkt 96.7 k, measured_height 96.0 k, measured_height_uom 96.0 k, usage_code 221, gml_name 125
New York (5)	bin 1.08 M, description 1.08 M, doitt_id 1.08 M, gml_name 1.08 M, source_id 1.08 M
Tokyo (20)	ade_xml 2.01 M, class_code 2.01 M, class_codespace 2.01 M, creation_date 2.01 M, description 2.01 M, measured_height 2.01 M, measured_height_uom 2.01 M, storeys_above_ground 2.01 M, storeys_below_ground 2.01 M, usage_code 2.01 M, usage_codespace 2.01 M, address_xml 1.81 M, 13_区市町村コード_大字_町コード_町_丁目コード (municipality–oaza–chome code)† 1.81 M, 大字_町コード (oaza–town code)† 1.81 M, 延べ面積換算係数 (total floor-area conversion factor)† 1.78 M, 町_丁目コード (town–chome code)† 1.72 M, 地区計画 (district plan)† 324.1 k, 説明注記 (descriptive note)† 184.1 k, 再開発等促進区を定める地区計画 (redevelopment-promotion district plan)† 6938, gml_name 4909

† The parenthetical English is our translation. The Japanese key is stored verbatim in the graph and is what a query must match.

Table 14: Question categories with template and instance counts, aggregated across all five cities.

Category	Capability probed	Tier	Tmpl.	Instances
Aggregate / filter / top-k	schema grounding, retrieval, ranking, incl. native per-city attributes	1	18	230
Multi-hop semantic	traversal over the compact schema, incl. native per-city attributes	1	11	105
Spatial	window, proximity, geometric multi-hop, 3D structure, density	1	12	462
Cross-source	(dis)agreement, coverage gaps, address completeness, dual roof shapes, OSM-only	1	13	237
Provenance-filtered	source and confidence constraints, provenance metadata	1	12	162
Coverage-aware	partial-coverage reasoning, incl. multi-material predictions	2	6	24
LoD3 showcase	facade/interior structure	2	4	21
Infeasible	hallucination resistance (guard-verified)	1	8	153
Multilingual	German/Japanese language variants; Japanese-key gold queries	1	3	95
Total			84	1394

Tier 2 categories are available only for Hamburg. *Multilingual* is an annotation applied across existing templates (German, Japanese, etc.) and is therefore excluded from the total.

Table 15: Authored difficulty rubric and its distribution over the 1394 frozen question instances. Conditions of lower levels may also apply at higher ones.

Level	Definition	Questions
1	one node label, one property, one aggregate; no filter	47 (3.4%)
2	one filter, one relationship hop, or ranking (`ORDER BY/LIMIT`)	272 (19.5%)
3	two composed elements: two hops, a grouped aggregate, one spatial predicate, one edge-property (provenance) filter, or a missing-value subtlety	714 (51.2%)
4	cross-concern composition: spatial × semantic, cross-source comparison with computation, coverage normalization, percentile or conditional share, schema bridging across languages	247 (17.7%)
5	three or more schema regions plus a geometric or set-level predicate: geometric multi-hop joins, per-entity grouped top-1 over two-hop patterns, n:m correspondence reasoning	114 (8.2%)

Table 16: Gold answer-type distribution over the 1394 frozen question instances.

Answer type	Questions
Scalar, numeric	556 (39.9%)
Table (multi-column or grouped)	352 (25.3%)
Ranked list (top-k, ordered and limited)	259 (18.6%)
Refusal (infeasible)	153 (11.0%)
Scalar, string	48 (3.4%)
List (single column)	26 (1.9%)

model is exposed to information unavailable to the other. Consequently, a model is expected to refuse requests for attributes that are absent from the provided schema text. Outputs are cached and evaluated offline. Generated Cypher queries are executed against the released graph and compared with gold answers after result-set canonicalization: row multisets are compared with column names removed and floating-point values rounded to 10^{-6}. As a result, adding an additional column or renaming an existing column alters the row signature and is therefore treated as incorrect rather than silently accepted.

The model qwen2.5-coder:7b is served locally through Ollama using the `Q4_K_M` quantization, a pinned model tag, greedy decoding (temperature = 0), the default context window, and a single generation per question. Claude Sonnet 5 is evaluated as a closed-book Claude Code subagent, with the model itself acting as the query generator rather than through the `ChatAnthropic` API endpoint. No `ANTHROPIC_API_KEY`, live database access, or external tools are available beyond the supplied schema text. The evaluated model

Table 17: Category × city feasibility. HH Hamburg, ZH Zurich, HEL Helsinki, TYO Tokyo, NYC New York.

Category	HH	ZH	HEL	TYO	NYC
Aggregate / filter / multi-hop	●	●	●	●	●
Spatial[1]	●	●	●	●	●
Cross-source / provenance[2]	●	●	●	●	●
Coverage-aware	●	–	–	–	–
LoD3 showcase[3]	●	–	–	–	–
Infeasible (other subfamilies)	●	●	●	●	●
Infeasible (absent layer)	–	●	●	●	●
Multilingual schema bridging	–	–	–	●	–
Dual roof-shape reporting	●	●	–	–	–
LoD3 confidence scalars[3]	●	–	–	–	–
Native single-city content[4]	–	●	●	●	–
OSM-only height[5]	●	●	●	●	●

[1] After metric-CRS reprojection at ingest for Tokyo and New York (Section 3.2). [2] After whole-city OSM fusion. [3] Applicable only to selected buildings in Hamburg (Section 3.5). [4] Single-home-city templates, gated by Section 3.2's requirement gate. Helsinki: floor area, storeys, volume, material, integration method, building status; Zurich: data vintage, revision history; Tokyo: basement storeys, town code, redevelopment district. Hamburg's own single-city content is the coverage and LoD3 rows above. [5] Cross-city by construction (no per-city gate needed), but only substantive where a real gap exists between the two sources' coverage, most pronounced for New York.

snapshot is `claude-sonnet-5`, tested on 2026-07-19 using the *high* reasoning-effort setting.

To improve token efficiency, questions within a split are submitted in batches of approximately 25 rather than as individual calls. Each batch prompt explicitly instructs the model to answer every question independently, relying only on the provided schema and rules and without influence from other questions in the batch. This design trades strict per-call isolation for a substantial reduction in the number of model calls. While answering many related questions in a single context may plausibly increase internal consistency relative to fully independent calls, this effect was not measured and no claim is made that the setup is equivalent to one-call-per-question evaluation.

H Representation-Learning Benchmark: Supplementary

Figure 6 shows the 2D t-SNE of Hamburg building embeddings from the provenance-agnostic and provenance-aware T3 matching encoders, the qualitative view referenced in Section 5.3. Table 18 reports the T2 roof-type cross-city transfer, limited to Hamburg and Helsinki, the only two cities that carry the label.

Table 18: T2 roof-type cross-city transfer, leave-one-city-out on the only valid two-city pair (macro-F1 on the merged 3-class label, mean over 3 seeds, best per row in bold).

Held-out (roof type)	agnostic GNN	aware GNN
Hamburg	0.203	**0.208**
Helsinki	**0.336**	**0.337**

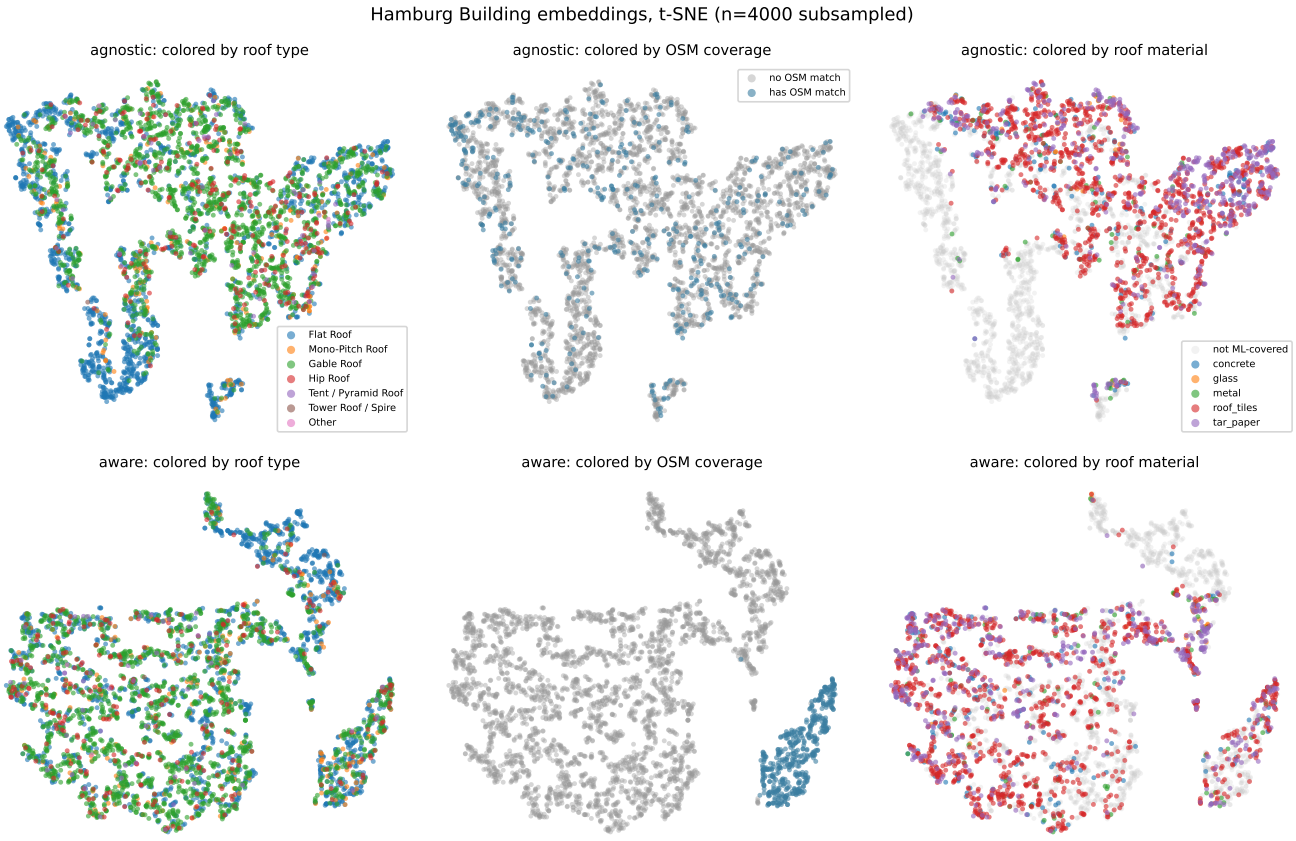


Figure 6: 2D t-SNE of Hamburg building embeddings from the T3 matching encoders. Columns show roof type, OSM coverage, and roof-material coverage, where gray marks buildings the ML layer never covered. Compared with the agnostic encoder, the aware encoder more clearly isolates OSM-covered buildings and yields a more structured space.