OpenCityCorpus: A Large-Scale, Harmonized, and LLM-Ready Corpus of Urban Data for Scientific Research

Junfeng Jiao[1], Sean Hardesty Lewis[1,2], Yiming Xu[1], Jihyung Park[1], Connor Phillips[1]
[1]The University of Texas at Austin
[2]Cornell University

Abstract

We propose *OpenCityCorpus*, an openly shareable, large-scale corpus that harmonizes public urban data from 200+ cities across Socrata, ArcGIS, and CKAN portals into a unified schema and an LLM-ready text representation. Fragmentation across municipal platforms has long impeded rigorous, cross-city science on climate, mobility, governance, and public health. Our dataset resolves schema heterogeneity, standardizes types and coordinate systems, and converts rows into semantically consistent factual statements, enabling retrieval-augmented generation, hypothesis testing, and transfer learning. The resource targets three AI-for-Science tasks: cross-domain scientific reasoning over coupled urban systems, surrogate modeling that complements physics-based simulators, and robust evaluation of tool-augmented LLM agents. We detail a feasible, privacy-preserving data-creation pathway, outline cost- and scale-aware operations for continuous refresh, and describe benchmarks designed to expose both the reach and the limits of current AI methods. By turning fragmented open portals into a single scientific substrate, *OpenCityCorpus* lowers barriers to high-impact, reproducible discovery.

1 Motivation and AI Task Definition

Urban systems exhibit strong interactions between engineered and natural processes. Progress in climate adaptation, transportation, public safety, and health equity is often bottlenecked by fragmented observational data rather than modeling ingenuity. *OpenCityCorpus* enables three tasks. First, it supports multi-domain scientific reasoning in which models must synthesize transit operations, emissions inventories, land use, permits, finance, and incident records to generate testable hypotheses across cities and years. Second, it provides supervision and conditioning signals for high-fidelity surrogate models that accelerate or augment physical simulators when spatial resolution, coverage, or boundary conditions are otherwise unavailable. Third, it establishes a retrieval substrate for LLM agents that must ground claims, trace evidence, and generalize under distribution shift, thereby revealing where tool use closes gaps and where reasoning still fails.

2 Dataset Rationale and Design

The core bottleneck in urban science is not the absence of data but the inability to use it across jurisdictions. Heterogeneous schemas, inconsistent metadata, and varying API conventions preclude systematic cross-city analysis despite the public nature of the sources. We harmonize approximately 200 GB of public datasets totaling over one billion records into a canonical, semantically typed schema with ISO 8601 time and WGS 84 geospatial standards. For LLMs and retrieval-heavy agents,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop:
Workshop on AI for Science: The Reach and Limits of AI for Scientific Discovery (NeurIPS 2025).

every normalized row is deterministically converted into a compact natural-language statement that preserves units, uncertainty flags, and provenance pointers. This dual representation of structured tables and aligned text creates a single collection where classical statistical models, graph learners, geospatial nets, and language models can share supervision signals and be evaluated on the same scientific questions.

3 Acceleration Potential for AI and Downstream Science

Access to harmonized, longitudinal, multi-city measurements changes both what is learnable and what can be falsified. For model development, the corpus exposes space–time holdouts and jurisdictional shift tests that are severe enough to probe extrapolation limits. For discovery, it makes counterfactual comparisons routine: for example, estimating whether service-frequency adjustments interact with heat events and equity metrics, or whether building-permit composition correlates with local energy use under policy changes. As a retrieval substrate, the text layer enables grounded question answering in which model claims can be traced to specific rows with units and timestamps, mitigating hallucinations [Lewis et al., 2020]. Together, these properties shorten the path from hypothesis to refutation by turning annotation-scarce domains into richly supervised ones, while also clarifying where current agents fail at cross-disciplinary reasoning.

4 Feasibility and Data-Creation Pathway

The pipeline follows an ELT design with a metadata registry of portals and endpoints, platform-specific connectors, a raw data lake, and automated harmonization. Schemas are inferred, mapped to a canonical ontology using string similarity and pattern templates, and standardized for types, time, and coordinates [Halevy et al., 2006]. The LLM-ready layer is produced by deterministic templates that render normalized records as factual statements with explicit provenance. Only public endpoints without login barriers are ingested, and rate limits are respected. Privacy risks are mitigated through k-anonymity and ℓ-diversity where quasi-identifiers exist, with differential privacy planned for small-area releases [Sweeney, 2002, Machanavajjhala et al., 2007, Dwork, 2008]. Licensing targets CC-BY 4.0 for aggregate releases; lineage metadata preserves original attributions and use terms.

5 Cost, Scalability, and Release Plan

At proposed scale, data egress and compute are modest relative to simulation-heavy efforts. A practical plan uses commodity cloud storage for the raw lake, columnar formats for the normalized tables, and nightly incremental refreshes; the template-rendered text is cached in shards for retrieval indices. A six-week initial build covers the largest portals, followed by rolling expansion to underrepresented regions. To maximize shareability, we will release versioned snapshots with fixed splits for benchmark stability, along with lightweight loaders for pandas, DuckDB, and common geospatial stacks. Governance includes a documented takedown process and a public issue tracker to capture schema bugs and equity concerns.

6 Evaluation and Expected Impact

Scientific value depends on what becomes falsifiable. The benchmark suite includes spatial and temporal generalization tests, retrieval-grounding audits with citation traceability, and ablations that quantify how much tool augmentation, including geocoders, unit normalizers, and graph-structured memory, actually improves LLM reasoning. Strong results would demonstrate reliable cross-city transfer and faithful retrieval, whereas failures will expose where language models remain brittle when faced with messy, real-world measurement streams. Either outcome advances the workshop's central goal: a clearer map of AI's reach and its limits in scientific discovery.

References

A. Halevy, A. Rajaraman, and J. Ordille. Data integration: The teenage years. *Proceedings of the VLDB Endowment*, 2006.

P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP. *NeurIPS*, 2020.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM TKDD*, 2007.

C. Dwork. Differential privacy: A survey of results. *TAMC*, 2008.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT networks. *EMNLP–IJCNLP*, 2019.

Appendix A: Legal and Ethical Notes

Only data that are publicly accessible without authentication are ingested. No technical barriers are circumvented, and rate limits are honored. While terms of service vary, our use targets scholarly research with clear attribution and opt-out pathways. We document lineage and apply privacy-preserving transforms prior to release.

Appendix B: Usage Examples

This appendix provides several illustrative, in-depth examples of how the harmonized and structured data within OpenCityCorpus can be utilized for advanced urban research and analysis. These tests are designed to showcase the core capabilities of the corpus in performing cross-jurisdictional analysis, enabling sophisticated data discovery, powering AI-driven question-answering, and facilitating data quality validation.

A.1. Cross-Jurisdictional Analysis: Residential vs. Commercial Construction

A primary advantage of OpenCityCorpus is the ability to perform nuanced comparative analyses across cities that use different data portal technologies and schemas. Without a harmonized corpus, such a query would require extensive, bespoke data wrangling for each city.

Objective: Compare the volume of new *residential* versus new *commercial* construction permits issued in the full year 2023 for Chicago, IL (Socrata) and Los Angeles, CA (ArcGIS Hub).

The Challenge of Heterogeneity: These two cities represent a common integration challenge. Their raw datasets use entirely different naming conventions and categorical values for the same concepts, as shown in Table 1.

Concept	Chicago (Socrata) field	Los Angeles (ArcGIS) field
Issue Date	`ISSUE_DATE`	`Issue_Date`
Permit Type	`PERMIT_TYPE`	`Permit_Sub-Type`
Occupancy	`reported_occupancy`	`Occupancy_Classification`
Status	`STATUS`	`permit_status`

Table 1: Illustrative example of unharmonized field names for building-permit datasets.

Methodology: Our harmonization pipeline resolves these inconsistencies. A single query was run against the OpenCityCorpus database using canonical field names.

- **Schema Mapping:** Fields like `ISSUE_DATE` and `Issue_Date` are mapped to a canonical ‘date_issued‘.
- **Value Standardization:** A dictionary-based mapping standardizes categorical values. For example, under the canonical ‘occupancy_type‘ field, Chicago’s ‘RESIDENTIAL‘ and Los Angeles’s ‘R-1‘ are both mapped to ‘residential‘. Similarly, ‘NEW CONSTRUCTION‘ and ‘New Const‘ are mapped to ‘new_construction‘.

The final query filters for records where ‘city‘ is in (’Chicago, IL’, ’Los Angeles, CA’), ‘date_issued‘ is within 2023, ‘permit_category‘ equals ’new_construction’, and then groups the results by the canonical ‘occupancy_type‘.

Results: The query returns a clean, comparable breakdown of construction activity, suitable for direct analysis (see **Table 2**).

City	Occupancy Type	Permit Count
Chicago, IL	Residential	9,876
Chicago, IL	Commercial	5,606
Los Angeles, CA	Residential	14,011
Los Angeles, CA	Commercial	7,734

Table 2: Illustrative harmonized comparison of new-construction permits issued in 2023.

Analysis: This harmonized result allows direct comparison across cities without extensive preprocessing. It could potentially enable urban economists and planners to immediately begin analyzing regional construction trends, housing development policies, and economic activity ratios without the weeks or months of data cleaning that would typically precede such work.

A.2. Advanced Semantic Data Discovery

Keyword-based search often fails to capture semantic intent, limiting dataset discovery. A traditional search for ”EV charging” might miss datasets titled ”Alternative Fueling Infrastructure” or permits for ”High-Amperage Electrical Work.” Our semantic search capability overcomes this by matching on conceptual meaning.

Methodology: Our semantic search is powered by generating sentence-level embeddings (using a Sentence-BERT model [Reimers and Gurevych [2019]]) for dataset titles, descriptions, metadata, and a sample of column names. User queries are also embedded into the same vector space, and we perform a cosine similarity search to find the most relevant datasets.

Example 1: Infrastructure Query

Natural Language Query: `"Public infrastructure for electric cars"`

Results:

1. *Dataset:* Public Electric Vehicle Charging Stations (**Score: 0.92**)
   - *City:* Austin, TX
2. *Dataset:* Alternative Fueling Corridors (**Score: 0.85**)
   - *City:* State of California
3. *Dataset:* New Electrical Service Permits - 2024 (**Score: 0.78**)
   - *City:* New York, NY
   - *Reasoning:* The system correctly inferred that permits for new electrical services are a proxy for the installation of new infrastructure like EV chargers.

Example 2: Social Policy Query

Natural Language Query: `"What after-school programs are available for teenagers in low-income neighborhoods?"`

Results:

1. *Dataset:* Parks and Recreation Program Catalog (**Score: 0.89**)
   - *City:* Phoenix, AZ
   - *Reasoning:* Contains program descriptions, age groups, and locations, allowing for joins with income data.
2. *Dataset:* Public Library Branch Events (**Score: 0.86**)
   - *City:* Philadelphia, PA
   - *Reasoning:* Captures teen-focused workshops and homework help sessions at library branches city-wide.
3. *Dataset:* Community Development Block Grants (**Score: 0.81**)
   - *City:* City of Boston, MA
   - *Reasoning:* Identifies non-profit organizations receiving city funding for youth services, often located in target neighborhoods.

These examples demonstrate how semantic search returns relevant datasets across different municipal departments, producing a holistic answer.

A.3. In-Depth Retrieval-Augmented Generation (RAG)

This test simulates a policy analyst using an LLM-powered chatbot that is connected to OpenCityCorpus via a RAG pipeline. The goal is to obtain a reliable, evidence-based summary without manually reading dense municipal budget documents.

Objective: Answer a nuanced comparative question about park funding, including specific project types and timelines.

User Question:

> *”Compare the recent funding for accessibility improvements at Franklin Park versus the Public Garden in Boston. Which projects are more recent, and are there any funds allocated for sensory-friendly play areas?”*

Retrieved Documents from OpenCityCorpus (Top 5):

1. **Snippet 1 (Source: Boston Capital Improvement Plan 2024-2028):** ”The project at Franklin Park with ID 789-A is for ’Pathway Repaving for ADA Compliance’ and received an allocation of $550,000 in fiscal year 2024.”
2. **Snippet 2 (Source: Boston Capital Improvement Plan 2022-2026):** ”The Public Garden received $85,000 in FY2023 for the installation of three new accessible water fountains and bench placements along the main lagoon path.”
3. **Snippet 3 (Source: Friends of the Public Garden - 2024 Annual Report):** ”A donation of $250,000 was secured in Q4 2024 for the restoration of the historic wrought-iron fence surrounding the Public Garden. The project prioritizes historical preservation and security.”
4. **Snippet 4 (Source: Boston Parks Dept. Public Meeting Minutes - May 2025):** ”Discussions were held regarding future capital projects. No funding has been allocated for sensory-friendly play areas in the current fiscal cycle, though it remains a community-requested item for future consideration.”
5. **Snippet 5 (Source: Boston Parks Dept. Grants 2023):** ”The ’Accessible Playgrounds Initiative’ grant was awarded to Christopher Columbus Waterfront Park for the amount of $120,000...”

LLM-Generated Answer (grounded by snippets):

> Here is a comparison of recent funding for accessibility improvements at Franklin Park and the Public Garden based on available data:
>
> **Franklin Park:**

> - A significant project for ’Pathway Repaving for ADA Compliance’ was allocated **$550,000** in fiscal year 2024. This is the most recent accessibility-focused project for this park in the data. (Source: Boston Capital Improvement Plan 2024-2028).
>
> **Public Garden:**
>
> - The park received **$85,000** in fiscal year 2023 for installing new accessible water fountains and benches. (Source: Boston Capital Improvement Plan 2022-2026).
> - A separate donation of $250,000 in late 2024 was for restoring the historic fence and was not designated for accessibility. (Source: Friends of the Public Garden - 2024 Annual Report).
>
> **Sensory-Friendly Play Areas:**
>
> - According to Parks Department meeting minutes from May 2025, **no funding has been allocated** for sensory-friendly play areas in the current fiscal cycle for any park, though it is a noted community request.
>
> In summary, Franklin Park has the more recent and larger single allocation for accessibility. The Public Garden’s accessibility funding was from the previous fiscal year.

This RAG interaction shows how the corpus enables the LLM to synthesize a precise, multi-part answer, correctly attributing funding, comparing timelines, and filtering out irrelevant information (the fence restoration and the grant for a different park).

A.4. Data Quality Anomaly Detection

Individual datasets often contain errors or outliers that are hard to detect in isolation. By aggregating data from over 200 cities, OpenCityCorpus creates a baseline distribution for common urban metrics, enabling powerful anomaly detection.

Objective: Detect anomalous 311 service request closure times in a specific city’s dataset by comparing it to the national distribution.

Methodology: We queried all ’pothole repair’ service requests from the 311 datasets for all cities in the corpus. We calculated the median time-to-close in days for each city. We then analyzed the distribution of these median times to identify statistical outliers.

Results: The analysis revealed that the distribution of median closure times across all cities is approximately log-normal, with a cross-city median of 7 days. However, the data for ’City X’ presented a stark anomaly.

- **Cross-City Median Closure Time:** 7 days
- **City X Median Closure Time:** 180 days
- **Further Anomaly in City X:** Analysis of raw records for City X showed that approximately 15% of pothole requests were marked as closed in ’0’ minutes.

Analysis: This outlier strongly suggests a systemic issue in City X’s data reporting rather than just slow service. The ’0 minute’ closures are likely data entry artifacts (e.g., automated closure on submission or a default value). The 180-day median could indicate a different definition of ’closure’ (e.g., end of the season) or a genuine backlog problem compounded by faulty data. This type of cross-jurisdictional sanity check, which immediately flags a dataset for quality review, is only possible with a large-scale, harmonized corpus like OpenCityCorpus.