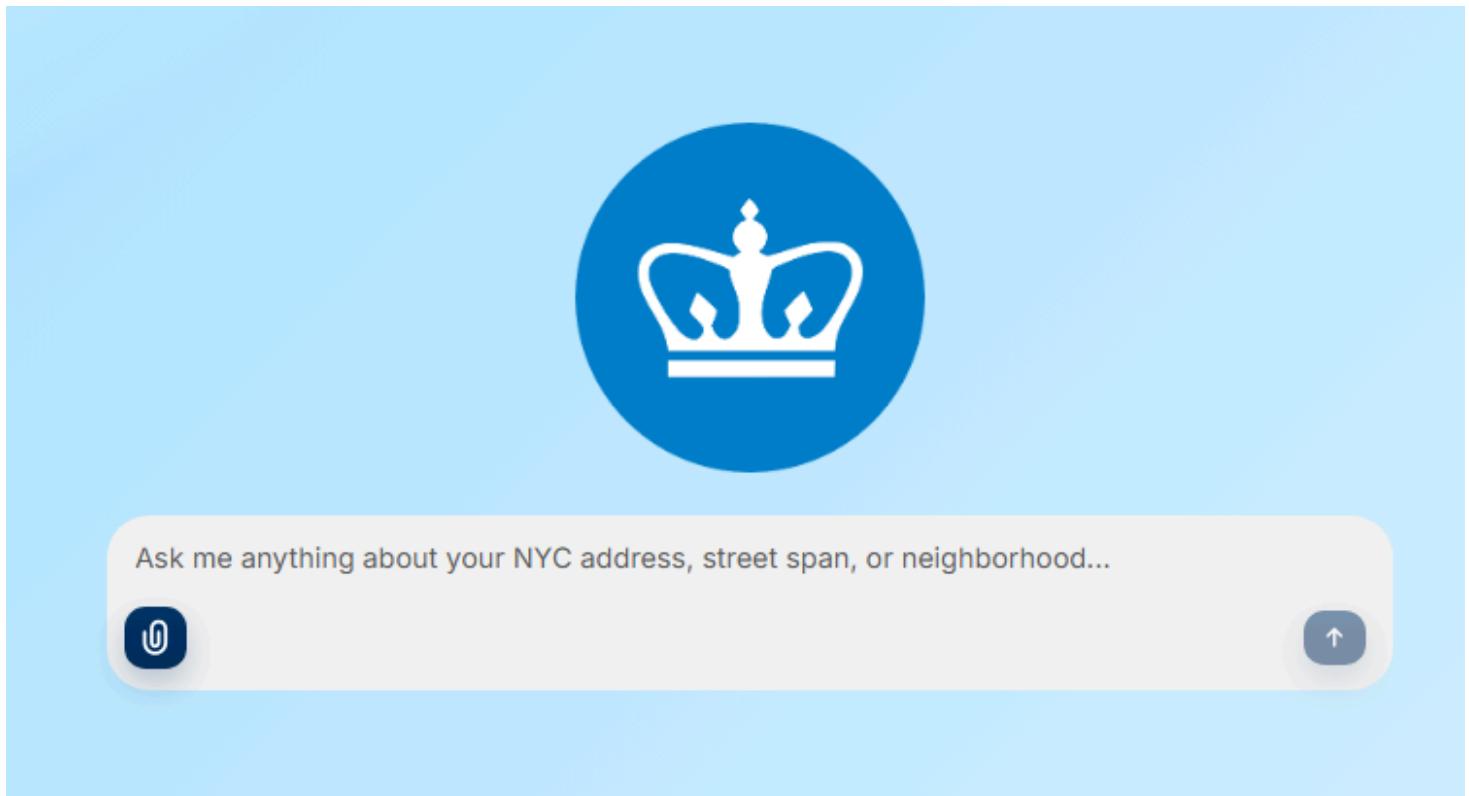


# Location-Aware GenAI-Powered Chatbot for Urban Risk Assessment

Columbia University, Data Science Institute

Fall 2025

Rishi Kumar (rk3326), Louis Chu (lc3920), Sharon Tsai (ct3263), Max Jan B Sebrechts (mjs2447), Kevin Ke (yk3108)



# Table of Contents

<b>Table of Contents.....</b>	<b>2</b>
<b>1. Problem Description.....</b>	<b>3</b>
1.1 Motivation and Background.....	3
1.2 Problem Statement.....	3
1.3 Project Overview.....	3
<b>2. Methods.....</b>	<b>4</b>
2.1 Primary Data Sources.....	4
2.1.1 Primary Land Use Tax Lot Output - Map (MapPLUTO) and LION.....	4
2.2 Geoclient API.....	5
<b>3. Exploratory Data Analysis.....</b>	<b>5</b>
3.1 Risk Category Definition.....	5
3.2 Geounit Exploration and Standardization.....	5
3.2.1 Choose BBL as the Common Spatial Unit.....	5
3.2.2 Building Geounit Adapters.....	5
3.2.3 From Geounits to GeoScope Mechanism.....	6
<b>4. LLM Integration and Workflow.....</b>	<b>6</b>
4.1 LLM Parser.....	8
4.2 LLM Risk Summarization.....	8
<b>5. System Usage, Evaluation, and Roadmap.....</b>	<b>8</b>
5.1 Example: Ambiguous Query (Condensed Output).....	8
5.2 Evaluation & Next Steps.....	9
<b>6. Conclusion.....</b>	<b>10</b>
<b>7. Contributions.....</b>	<b>11</b>
<b>8. Appendix.....</b>	<b>11</b>
<b>9. Acknowledgements.....</b>	<b>17</b>
<b>10. References.....</b>	<b>18</b>

# 1. Problem Description

This Capstone project is a collaboration between **Columbia University** and **Town+Gown: NYC @ DDC**.

## 1.1 Motivation and Background

Public infrastructure and building projects in New York City must comply with a variety of environmental, zoning, and safety regulations. However, the information needed to assess project risk is scattered across dozens of publicly available datasets maintained by different agencies - each with its own formats, geospatial units, and update schedules.

Project managers at the NYC Department of Design and Construction (DDC) currently spend significant time manually cross-referencing data from sources such as MAPPLUTO, LION, environmental datasets, crime datasets, and many others to evaluate whether a proposed project site lies within areas of concern (e.g., asbestos contamination, flood risk, etc.). This manual, spreadsheet-driven process slows down project reviews and increases the likelihood of human error, especially when multiple datasets use inconsistent geospatial reference systems.

## 1.2 Problem Statement

The core problem this project addresses is **the fragmentation and complexity of geospatial risk data used by NYC project managers**. Each dataset is indexed by different geounits, such as Borough-Block-Lot (BBL), Building Identification Number (BIN), Street Segment (LION), Neighborhood Tabulation Area (NTA), and Precinct, among others, making it challenging to automatically align information across data sources to relevant project sites.

Our goal is to **develop a GenAI-powered, location-aware chatbot** that can interpret user input (e.g., a question about an address), convert it into the appropriate geounits, retrieve relevant risk datasets, and summarize the key findings in clear, project manager-friendly language, all while maintaining an ongoing conversation with the user without prior expertise in geospatial analysis. However, the Minimum Viable Product (MVP) version of this project focuses solely on asbestos risks, demonstrating the end-to-end workflow for user query, geospatial normalization, dataset retrieval, risk summarization, and chatbot response within an initially limited domain.

## 1.3 Project Overview

Recent research has highlighted the growing potential of chatbots and large language models (LLMs) for context-aware reasoning. Zhu *et al.* [1] showed that chatbot-based hazard training can improve safety awareness in construction settings, while Polo-Rodríguez *et al.* [2] demonstrated how LLM-driven systems can use location context to deliver adaptive, real-time interactions. Building on these advances, our project extends such capabilities to the urban infrastructure domain through a **location-aware, GenAI-powered chatbot** that synthesizes geospatial data into actionable risk insights for New York City project managers.

The chatbot functions as a **geospatial intelligence assistant**, allowing project managers and city officials to ask natural-language questions such as “*What environmental risks are near 237 Park Avenue?*” The system processes each query through a sequence of modular operations in its **data-to-dialogue pipeline**, producing concise, evidence-based summaries. The workflow comprises five key components that ensure automation, interpretability, and scalability.

1. **Data Ingestion & Preprocessing:** The system integrates data covering various urban risks. All datasets are standardized to a common **Borough-Block-Lot (BBL)** unit using geometric joins, the **NYC Geoclient API**, and conversion adapters. This enables consistent spatial queries and seamless interoperability across data sources.

2. **Geospatial Normalization & GeoScope Generation:** When a user provides a location, the Geoclient API retrieves its identifiers (BBL, BIN, latitude/longitude). The **GeoScope module** defines spatial context using street-span tracing via **LION geometry** for the chatbot to assess both the target lot and its surrounding area for contextual risk analysis.
3. **Query Interpretation & Dataset Selection:** An **LLM Parser** interprets user intent, extracts structured address components, and classifies queries into one or more risk categories (e.g., *Environmental & Health Risks, Construction & Permitting*). These categories are mapped to relevant datasets through a configurable routing dictionary, ensuring efficient retrieval and coherent summarization.
4. **Risk Retrieval & Summarization:** The **DataHandler** loads and filters datasets based on GeoScope output, merging relevant records into a unified view. A specialized **Risk Summarization LLM** analyzes this data to produce a concise, natural-language summary of key findings—such as asbestos filings, permit statuses, and material types—while distinguishing between active, historical, and upcoming risks.
5. **Conversational Interface & Iterative Dialogue:** The **Conversational LLM** manages dialogue flow and context, enabling follow-up queries like “include nearby lots” or “compare with another address.” This modular design ensures transparency, allowing users to preview underlying data and expand to new risk domains without re-engineering the system.

Together, these components form an **end-to-end modular pipeline** that converts unstructured user queries into structured geospatial intelligence and back into clear, context-aware responses. The architecture combines **deterministic geospatial computation** (e.g., address parsing, GeoScope, spatial joins) with **generative reasoning** (LLM-based parsing and summarization), ensuring outputs that are both factually grounded and conversationally meaningful. Ultimately, the framework establishes a **scalable foundation** for integrating additional datasets and expanding into multi-domain urban risk assessment.

## 2. Methods

### 2.1 Primary Data Sources

All datasets used in this project were sourced from NYC Open Data, the City of New York’s official website for providing public information across domains such as zoning, infrastructure, and environment. Of the 20+ datasets identified (see Appendix A1 for further details on each dataset), MapPLUTO and LION form the project’s geospatial framework, while the remaining datasets serve as domain-specific data for risk insights. MapPLUTO provides Borough-Block-Lot (BBL), a unique identifier used in NYC to locate and reference individual tax lots that serve as the primary spatial key. LION contains street segment geometries that link intersection-based queries to nearby BBLs. The NYC Geoclient API complements them by geocoding addresses and returning related BBL and street information. Because datasets reference different geounits (e.g., BINs, NTAs, precincts, coordinates), all were standardized to the BBL level using MapPLUTO. Additional sources, such as LION and NTA2020, were required to map street spans and Neighborhood Tabulation Areas, which Geoclient does not provide. This ensured consistent spatial alignment before integration into the chatbot.

#### 2.1.1 Primary Land Use Tax Lot Output - Map (MapPLUTO) and LION

The MapPLUTO dataset, maintained by the NYC Department of City Planning, combines detailed tax lot information with land-use, zoning, and building data. It provides parcel-level geometries across the five boroughs of NYC using BBL alongside other geospatial identifiers. MapPLUTO served as one of the primary reference layers, as it contains BBL polygons that could be overlapped with geometries from all other datasets. To optimize performance, the dataset was loaded using Python’s `lru_cache` decorator, which enabled efficient memory management while maintaining fast access to lot geometries. The LION dataset shared similar functionality to the MapPLUTO dataset. Although it doesn’t include a BBL identifier, it covers public infrastructure, such as roads and utilities, using street and borough codes. In this project, LION was used to link

street-level datasets (e.g., construction and traffic data) to their corresponding BBLs by matching spans and intersections. The dataset was read and processed similarly, using `lru_cache` to maintain the same memory-efficient strategy.

## 2.2 Geoclient API

Once the datasets are standardized using MapPLUTO and LION, the NYC Geoclient API was integrated to handle natural-language user queries. When a user enters a location prompt (e.g., “Summarize the asbestos risk in 237 Park Ave.”), the LLM-parsed input is fed to the Geoclient API to retrieve attributes such as BBL, BINs, longitude, and latitude. The retrieved BBL will then be merged and used to search across all datasets for relevant information. This approach ensured that all datasets, regardless of their original geospatial unit, can be queried and integrated consistently through a unified geospatial framework.

# 3. Exploratory Data Analysis

## 3.1 Risk Category Definition

During the exploratory phase, our primary objective was to understand the scope, spatial resolution, and thematic relevance of available NYC Open Data sources. After examining over 20 datasets, we organized them into six major **risk categories—Environmental & Health Risks, Zoning & Land Use, Construction & Permitting, Transportation & Traffic, Public Safety & Social Context, and Comparative Site Queries** (see details in Appendix A4). These categories capture the main types of questions city agencies and project managers are likely to encounter.

Each dataset may belong to multiple categories. For example, the *Asbestos Control Program* dataset contributes to both **Environmental & Health Risks** and **Construction & Permitting**, since asbestos filings are both environmental and construction-related. This tagging process produced a configuration mapping (`cat_to_ds`) that now enables the LLM query router to automatically match user intent (e.g., “environmental risks”) to the correct datasets.

## 3.2 Geounit Exploration and Standardization

In the previous section, we discussed that different NYC Open Data sources describe location using different geographic units (geounits). To address this challenge, we systematically examined the geounits in each dataset and developed a conversion framework to unify them into a standard reference unit.

### 3.2.1 Choose BBL as the Common Spatial Unit

The **Borough–Block–Lot (BBL)** is the most fundamental parcel identifier used across New York City’s property, zoning, and environmental datasets. We selected BBL as our *base geounit* because it offers three key advantages. First, it is widely used across critical datasets, including MAPPLUTO, Department of Buildings (DOB) permits, and zoning maps, making it a universal reference point for geospatial alignment. Second, it provides fine-grained parcel-level precision, ideal for construction and land-use risk analysis. Finally, other spatial identifiers—such as BINs, precincts, or neighborhood tabulation areas—can be reliably mapped to BBL through geometric or API-based transformations (e.g., the NYC Geoclient API or polygon joins). By unifying all datasets at the BBL level, we ensured that data from different sources could be cross-joined, aggregated, and visualized consistently in downstream analyses.

### 3.2.2 Building Geounit Adapters

The key geounits we encountered and integrated are provided in Appendix A3: Geounits. To translate between these spatial references, we developed modular “adapter” scripts. Each adapter converts between BBL and another geounit type (bidirectional), allowing flexible filtering and merging across datasets. Each adapter can be called independently, but all

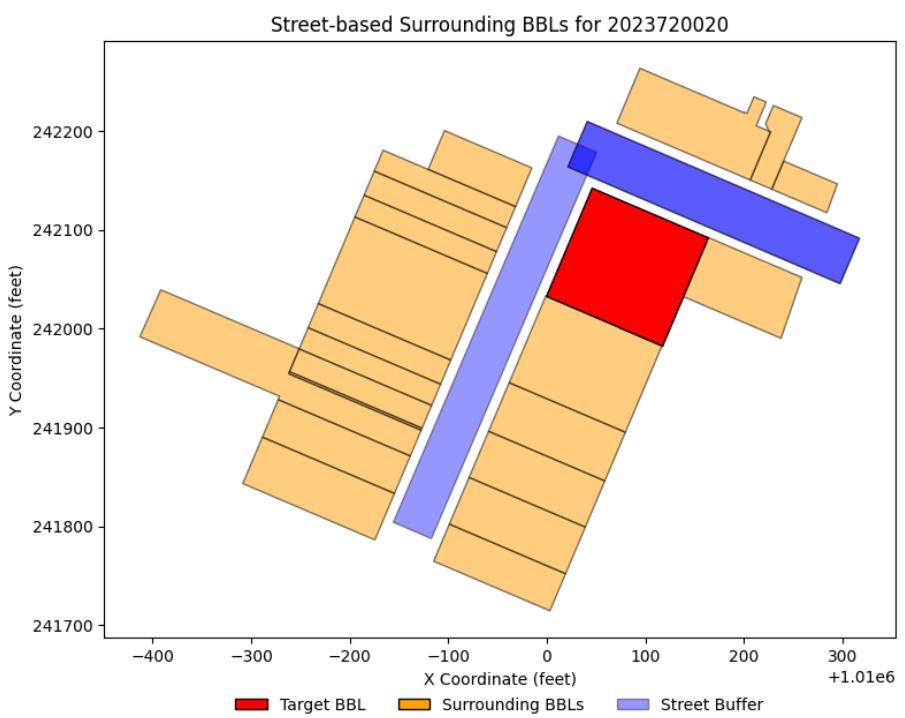
follow a consistent API, allowing the system to automatically pick the correct conversion route based on a dataset's metadata (e.g., whether it uses NTA, BIN, or raw coordinates).

### 3.2.3 From Geounits to GeoScope Mechanism

While BBL standardization allows all datasets to be compared at a parcel level, risk evaluation often requires analyzing context — not just a single lot, but also its surrounding environment. To achieve this, we implemented a *GeoScope mechanism* that automatically determines the spatial scope of each user query. GeoScope starts with a user's address (converted to BBL via Geoclient) and finds surrounding parcels via spatial buffering along the street network (using LION). When GeoScope determines a set of relevant BBLs for a user query, the system checks each dataset's configuration (GeoConfig) to identify what spatial unit it uses. Then, it automatically applies the corresponding adapter to translate the BBL-based scope into the dataset's native unit before applying filters.

Given a user input like “*Are there asbestos filings near 237 Park Avenue in Manhattan?*”, GeoScope performs three key steps:

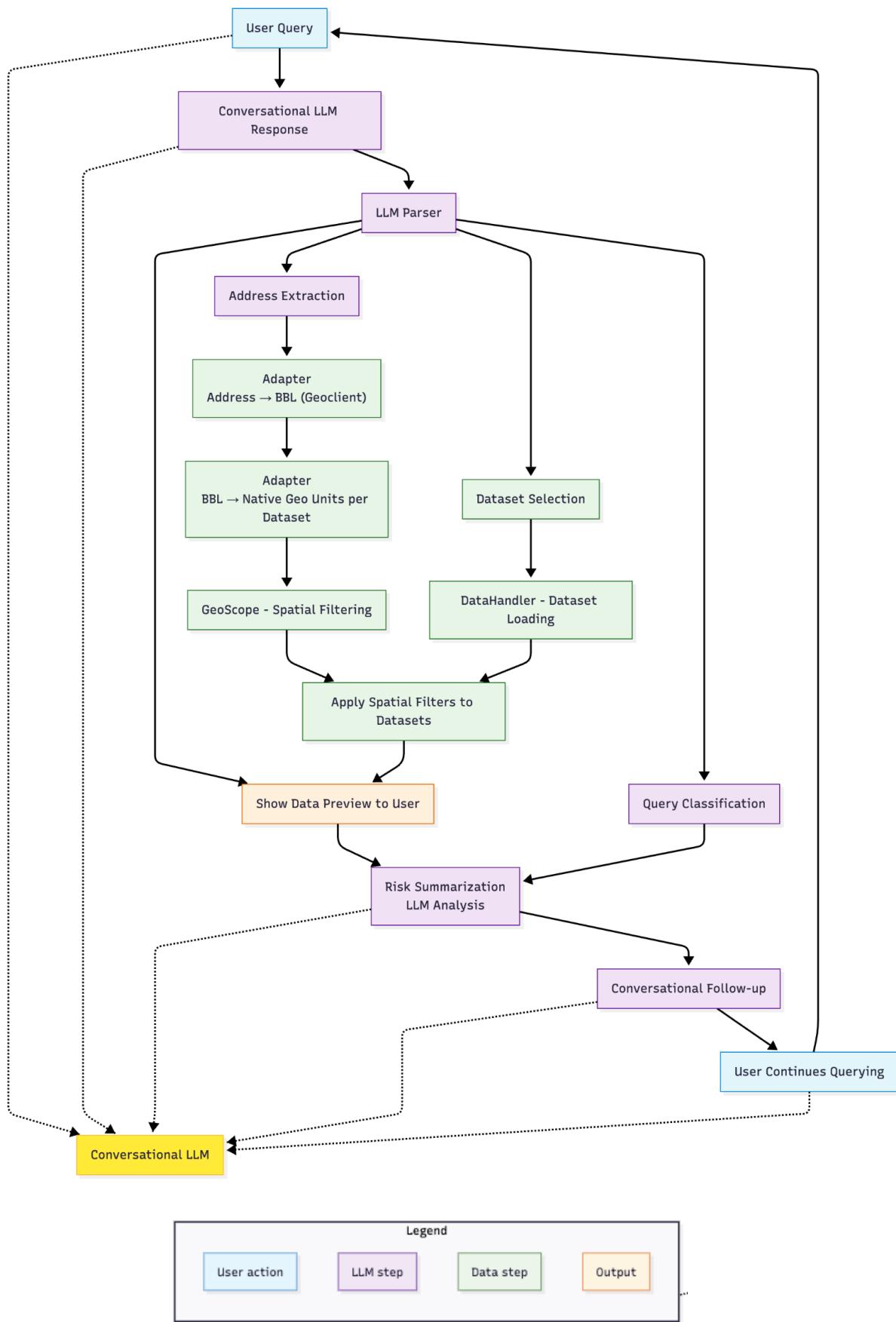
1. **Address → BBL:** The system first calls the NYC Geoclient API to convert the address into a standardized BBL.
2. **BBL → Surrounding BBLs:** Using LION street geometry and MAPPLUTO, it finds all nearby parcels that are spatially related to the input BBL.
3. **Scope → Dataset Filter:** The surrounding BBLs are then translated into each dataset's native geounit (e.g., coordinates, precincts) using the appropriate *geounit adapter*. These converted spatial units are used to filter relevant records (e.g., asbestos filings, flood zones, or permits) before risk summarization.



## 4. LLM Integration and Workflow

In the application, there is one “conversational LLM,” which can be seen as the chatbot the user is interacting with. This stateful, conversational LLM pulls in relevant inputs from the various stages and generates appropriate responses and follow-ups. At multiple stages throughout the pipeline (i.e., LLM Parser, Risk Summarization), individual, stateless LLM calls are also performed to complete various tasks. All LLM interactions are made through calls to the Gemini API, funded by the GCP credits we received for this course.

## End-to-End System Workflow for Location-Aware Chatbot



The application's flow starts with the **user's query**. The conversational LLM provides an initial response, and the **LLM parser extracts the address, relevant categories, and datasets**. Based on the extracted address and categories, the **GeoUnit adaptors and GeoScope mechanisms extract information pertinent to the query from the selected datasets using the DataHandler**. A preview of the data is then already shown to the user, after which the **LLM risk summary analysis initiates**. This risk summary is then shown to the user, **answering the original query**, after which the conversational **LLM generates a follow-up response**. This response provides additional insights and recommendations for further queries. Finally, the user can ask another query, which restarts the whole pipeline with a new input. Importantly, after this new query, the conversational context of the previous query is preserved, and chat history is persisted throughout the user session. The next subsections describe the two LLM components central to this workflow: the Parser and the Risk Summarization module.

## 4.1 LLM Parser

The LLM Parser achieves three tasks: address extraction, query classification, and dataset selection. First, from the unstructured user query, an LLM extracts all the mentioned addresses in a structured format. This was done by writing a comprehensive prompt that was iteratively refined to contain various instructions to improve the accuracy of the generated output. Furthermore, a few-shot prompting strategy is employed, adding multiple correct input-output pairs to the prompt, thereby greatly improving the LLM's response accuracy. The LLM is then asked to return the address in a structured JSON format, enabling easy parsing of its output. Lastly, if the LLM fails, a fallback regex extraction is used.

Secondly, the LLM parser also classifies the query into one of 6 categories as described in section 3.1, and expanded upon in Appendix A5. This is again done using a comprehensive system prompt and in a few-shot regime. Furthermore, for this task, we ask the LLM to give a confidence score for each selected category. This was inspired by Tian *et al.* [3], who showed that verbalized confidence scores as output tokens can improve the LLM's accuracy. Thirdly, the LLM's chosen categories are used to select datasets to query, mapping each category to a set of relevant datasets. This ensures that only information pertinent to the query is retrieved, rather than retrieving data from all available datasets.

## 4.2 LLM Risk Summarization

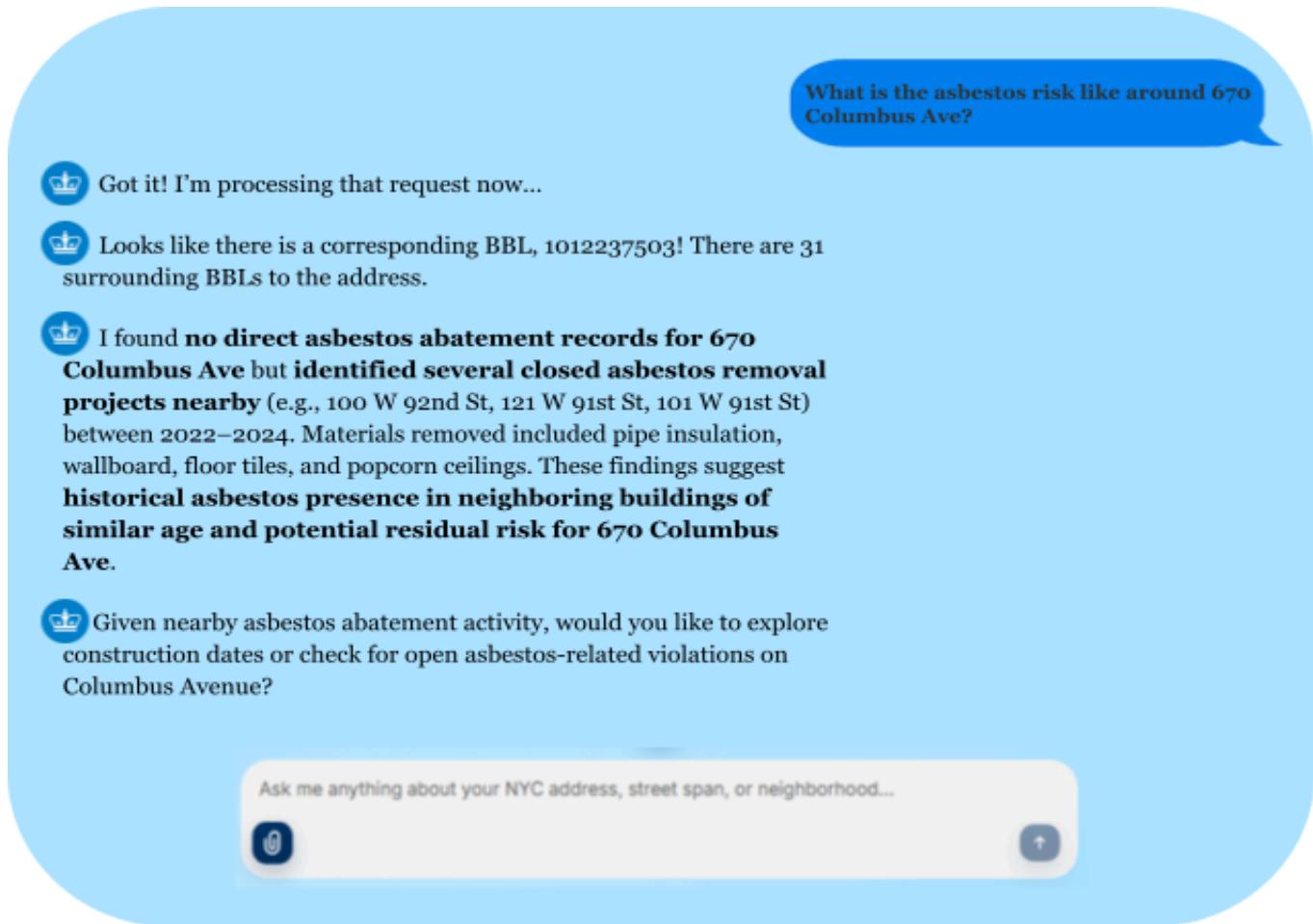
An LLM analyzes the retrieved data in the context of the user's query and generates a comprehensive risk assessment summary. This is a crucial part of the pipeline, as it brings everything together to answer the user's question. It is achieved by creating a descriptive prompt that contains all relevant information (extracted data, user query, addresses, etc.) in a structured manner, and feeding it to a separate LLM tasked with generating a risk summary. The extracted rows from the relevant datasets are given in markdown format, allowing the LLM to parse the information easily. Furthermore, through significant prompt engineering efforts, a lengthy explanation was created that specifies precisely what the LLM needs to do and which information can be helpful for this task. This includes, but is not limited to, a description of each dataset from which data was extracted, and general context about the role of the LLM in assisting with risk and compliance questions in building and construction projects in NYC. In doing so, the system transforms raw geospatial data into actionable, human intelligence.

# 5. System Usage, Evaluation, and Roadmap

## 5.1 Example: Ambiguous Query (Condensed Output)

The example below shows the chatbot's **ability to handle complex, ambiguous user input and synthesize a nuanced, context-aware response**. When asked, “*What is the asbestos risk like around 670 Columbus Ave?*”, the system interprets the vague spatial intent (“around”), locates the nearest geounits through the Geoclient API using only a partial input (house number + street), and identifies relevant nearby asbestos abatement records even when no direct matches exist for the queried

address. The generated summary contextualizes these findings—highlighting historical asbestos activity, proximity-based



residual risk, and relevant neighboring projects—while suggesting meaningful follow-up questions to continue the dialogue.

While this example highlights **advanced reasoning and spatial nuance**, the chatbot also performs efficiently on simpler, direct queries—such as “*Is there asbestos risk at 1633 Broadway?*”—where it can extract address details, locate a precise BBL match, and summarize site-specific information like material type, project status, and timeline with minimal ambiguity. Moreover, in comparative queries such as “*Compare asbestos risk between 405 Lexington Avenue and 1633 Broadway,*” the chatbot can correctly parse multi-address intent and return structured summaries. Together, these capabilities illustrate **the chatbot’s flexibility across varying levels of query complexity, from direct address lookups to sophisticated, spatially aware reasoning**. The full LLM-generated output for the shown interaction is provided in Appendix A1.

## 5.2 Evaluation & Next Steps

We evaluate the MVP’s pipeline using a structured methodology against a Ground Truth Set of 20 diverse queries (see Appendix A5). We manually verified the correct answer for each to create an answer key.

We then scored the chatbot’s performance against this set on five key metrics. **Parsed Address Accuracy** measures the percentage of queries where the system correctly parsed the user’s intent. **BBL Accuracy** measures the percentage of queries that were successfully matched to the correct BBL. **Record Match Accuracy** measures the percentage of queries that retrieved the correctly scoped data. **Risk Summary Faithfulness** is a qualitative Pass/Fail review for all 20 queries, ensuring

the LLM's summary accurately reflects the retrieved data with no hallucinations. Finally, **Conversational Quality** is a subjective 1-5 average rating of the helpfulness of the bot's follow-up question.

Evaluation Metrics	System Component Tested	Score
1. Parsed Address Accuracy	Prompt Interpreter	100% (20/20)
2. BBL Accuracy	Adapters	100% (20/20)
3. Record Match Accuracy	Geoscope / LLM Parser / DataHandler	100% (20/20)
4. Risk Summary Faithfulness	Risk Summarizer (LLM)	90% (18/20)
5. Conversational Quality	Conversational Aspect (LLM)	4.3 / 5.0 (Average over 20)

Minor inaccuracies in risk summaries and conversational follow-ups reflect the inherent probabilistic nature of LLMs, which can generate unsupported statements without explicit grounding. Our findings validate the robustness of the MVP while highlighting key areas for improvement—particularly in guiding the LLM's reasoning and expanding the system's scope.

While our MVP has validated the chatbot's core query and summarization pipeline, the next phase focuses on improving interpretive consistency to mitigate LLM limitations and expanding domain coverage. Specifically, we will:

1. **Integrate Deterministic Summary Guardrails:** Implement a deterministic summary framework as a *guidance layer* to steer the LLM toward well-defined risk insights. Rather than correcting the LLM, this framework pre-computes structured indicators particular to each domain to highlight what truly matters. These insights act as expert cues, reducing hallucination and promoting consistent, domain-relevant summaries.
2. **Expand Multi-Domain Knowledge Base:** Extend the system beyond asbestos to include datasets on other risks (see Appendix A2), enabling cross-domain risk analysis within a single conversational framework.

## 6. Conclusion

This project demonstrates the feasibility and value of a **location-aware, GenAI-powered chatbot** for urban risk assessment in New York City. By integrating geospatial datasets such as **MAPPLUTO**, **LION**, and the **NYC Geoclient API** into a unified **Borough-Block-Lot (BBL)** framework, the system enables seamless alignment of heterogeneous public datasets for automated spatial reasoning. The **Minimum Viable Product (MVP)**, focused on asbestos risk, provides an end-to-end proof-of-concept spanning natural language query interpretation, geospatial standardization, dataset retrieval, and LLM-based summarization—delivering clear, evidence-based insights in conversational form.

Initial evaluation results show that the chatbot performs strongly in **address parsing, contextual summarization, and accuracy of retrieved risk information**, achieving reliable, interpretable outputs within reasonable latency. The hybrid architecture—combining deterministic geospatial computation with generative reasoning—proves effective for converting unstructured user queries into structured, actionable intelligence.

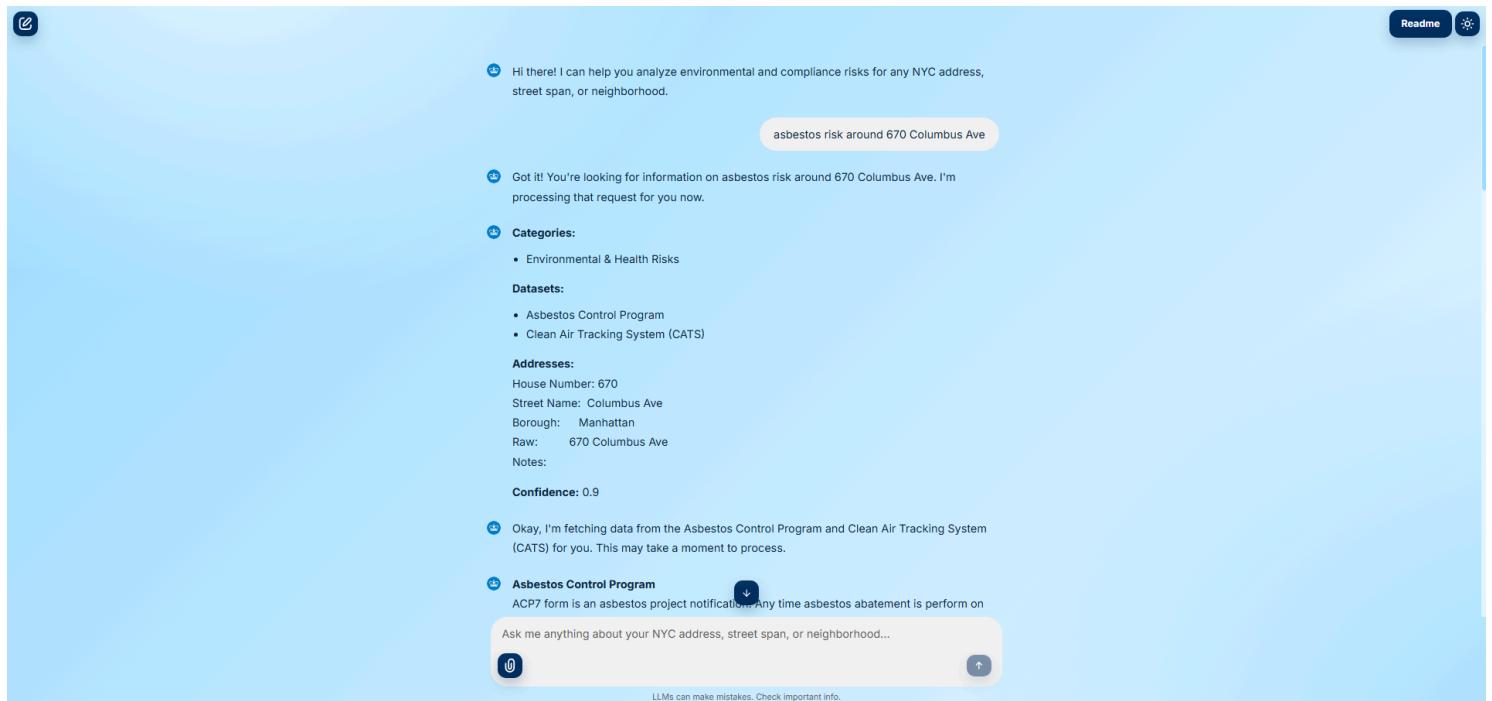
Beyond the MVP, this framework establishes a **scalable foundation** for multi-domain expansion, including environmental, infrastructure, and regulatory risk assessments. With further dataset integration and enhancements to deterministic summarization and multi-address comparison, the system can evolve into a comprehensive **decision-support tool** for NYC agencies. Ultimately, this Capstone project demonstrates how generative AI, grounded in precise geospatial computation, can redefine how complex urban data is accessed, interpreted, and communicated—making risk assessment faster, more consistent, and more transparent across the city's built environment.

## 7. Contributions

Each team member played a key role in designing and implementing the MVP. All five members collaboratively performed exploratory data analysis across more than 20 NYC Open Data sources and co-developed the bidirectional geounit adapters for geospatial standardization. Louis led the development of the LLM Prompt Interpreter and the DataHandler module, overseeing dataset retrieval and integration. Kevin designed and implemented the GeoScope mechanism that automates spatial context generation and buffering. Max developed the Risk Summarization pipeline, creating structured prompts and conducting extensive prompt engineering for the LLM-based analysis. Sharon built the Deterministic Summarization framework, introducing rule-based risk scoring to ground LLM outputs in factual data. Rishi designed the overall system architecture, orchestrating the multi-stage pipeline, UI design, and integration of the conversational interface that connects all components into a cohesive end-to-end chatbot system.

## 8. Appendix

### A1: Full Ambiguous Query Chatbot Conversation



**ACP7**

ACP7 form is an asbestos project notification. Any time asbestos abatement is performed on quantities greater than a minor project amount, the applicant has to file this form with DEP Asbestos Control Program (ACP). All asbestos documents are filed through the Asbestos Reporting and Tracking System (ARTS) E-file system. This system is web based and entirely paperless. All information on the ACP7 is essential to meet the requirements set forth in the asbestos rules and regulations Title 15, Chapter 1 (RCNY). ACP enforcement staff utilizes this form for inspection of asbestos abatement activities.

Preview:

tru	start_date	end_date	status_description	street_activity
TRU1849MN22	2022-07-15T00:00:00.000	2023-07-21T00:00:00.000	Closed	No
TRU0120MN23	2023-03-20T00:00:00.000	2024-01-19T00:00:00.000	Closed	No
TRU1374MN23	2023-10-05T00:00:00.000	2024-06-07T00:00:00.000	Closed	No
TRU1374MN23	2023-10-05T00:00:00.000	2024-06-07T00:00:00.000	Closed	No
TRU2897MN23	2024-01-24T00:00:00.000	2024-11-19T00:00:00.000	Closed	No

**Clean Air Tracking System (CATS)**

Ask me anything about your NYC address, street span, or neighborhood...

LLMs can make mistakes. Check important info.

**CATS**

TRU2897MN23	2024-01-24T00:00:00.000	2024-11-19T00:00:00.000	Closed	No
-------------	-------------------------	-------------------------	--------	----

**Clean Air Tracking System (CATS)**

Preview:

requestid	applicationid	requesttype	house	street	borough
54696	CA059089	REGISTRATION	142	FULTON STREET	MANHATTAN
378827	CA022047	CERTIFICATE TO OPERATE	70	nan	MANHATTAN
394994	CR081417	REGISTRATION	1581	FULTON AVENUE	BRONX
118102	CA026270	CERTIFICATE TO OPERATE	2419	DAVIDSON AVENUE	BRONX
456474	CA062298	REGISTRATION	19-47	80 STREET	QUEENS

Ask me anything about your NYC address, street span, or neighborhood...

LLMs can make mistakes. Check important info.

**Risk Summary:**  
Based on the data, there are no specific asbestos abatement projects listed directly for 670 Columbus Ave.

However, the "Asbestos Control Program" dataset shows multiple **closed** asbestos abatement projects in the immediate vicinity:

- **100 West 92nd Street** (very close to your requested address) had numerous abatement projects between March 2023 and August 2024. These projects involved the removal of various asbestos-containing materials, including:
  - Joint compound on walls
  - Vertical pipe insulation
  - Wallboard/joint compound
  - Enclosed vertical pipe insulation
  - Floor tiles (including VAT - Vinyl Asbestos Tile)
  - Popcorn ceilings
- **121 West 91 Street** (Trinity School, one block south) had a closed project in 2022-2023 for the removal of "ACM debris / infill."
- **101 West 91 Street** (also listed as Trinity School, potentially part of the same complex as 100 W 92nd St due to BIN) had a closed project in 2022-2023 for the removal of "TSI Pipe Insulation."

The fact that several asbestos removal projects have been completed in such close proximity indicates that asbestos-containing materials were present in buildings in this area. While all these listed projects are "Closed," meaning the abatement work was completed, it suggests that buildings of similar age and construction in the neighborhood, including potentially 670 Columbus Ave, may have historically contained or currently contain asbestos that has not yet been identified or abated.

 Given the recent asbestos abatement activity in nearby buildings, would you be interested in exploring the construction dates of properties on Columbus Avenue to infer potential asbestos presence, or perhaps looking into any open violations related to asbestos in the area?

Ask me anything about your NYC address, street span, or neighborhood...  

LLMs can make mistakes. Check important info. 

abatement projects in the immediate vicinity:

- **100 West 92nd Street** (very close to your requested address) had numerous abatement projects between March 2023 and August 2024. These projects involved the removal of various asbestos-containing materials, including:
  - Joint compound on walls
  - Vertical pipe insulation
  - Wallboard/joint compound
  - Enclosed vertical pipe insulation
  - Floor tiles (including VAT - Vinyl Asbestos Tile)
  - Popcorn ceilings
- **121 West 91 Street** (Trinity School, one block south) had a closed project in 2022-2023 for the removal of "ACM debris / infill."
- **101 West 91 Street** (also listed as Trinity School, potentially part of the same complex as 100 W 92nd St due to BIN) had a closed project in 2022-2023 for the removal of "TSI Pipe Insulation."

The fact that several asbestos removal projects have been completed in such close proximity indicates that asbestos-containing materials were present in buildings in this area. While all these listed projects are "Closed," meaning the abatement work was completed, it suggests that buildings of similar age and construction in the neighborhood, including potentially 670 Columbus Ave, may have historically contained or currently contain asbestos that has not yet been identified or abated.

 Given the recent asbestos abatement activity in nearby buildings, would you be interested in exploring the construction dates of properties on Columbus Avenue to infer potential asbestos presence, or perhaps looking into any open violations related to asbestos in the area?

Ask me anything about your NYC address, street span, or neighborhood...  

LLMs can make mistakes. Check important info. 

## A2: Data Descriptions

Dataset Name	Spatial Unit	Risk Category	Description
NYC OpenData Zoning and Tax Lot Database	BBL	Zoning & Land Use	Contains zoning and land use classifications for each tax lot, used to determine regulatory and development constraints.

NYC OpenData PLUTO	BBL	Zoning & Land Use, Public Safety & Social Context	Provides detailed building and land use data, including lot area, ownership, and construction year.
NYC OpenData Automated Traffic Volume Counts	Lat/Lon, Street	Public Safety & Social Context, Transportation & Traffic	Records traffic volume at bridges and major roads, supporting transportation and congestion analysis.
NYC OpenData Motor Vehicle Collisions	Lat/Lon, Street	Transportation & Traffic, Public Safety & Social Context	Documents traffic incidents and crash locations, used to identify transportation risk areas.
DOB Permits	BBL	Construction & Permitting	Permits for construction and demolition activities in the City of New York. Each record represents the life cycle of one permit for one work type.
Asbestos Control Program	BBL	Environmental & Health Risks	Any time asbestos abatement is performed on quantities greater than a minor project amount, the applicant has to file this form with DEP Asbestos Control Program (ACP).
Digital City Map Shapefile	Lat/Lon, EPSG:4326	Zoning & Land Use	The Digital City Map (DCM) data represents street lines and other features shown on the City Map, which is the official street map of the City of New York.
Historic Districts Map	Polygon	Zoning & Land Use	Defines boundaries of designated historic districts.
LION	EPSG:2263	Zoning & Land Use	A single line street base map representing the city's streets and other linear geographic features.
Zoning GIS Data	EPSG:2263	Zoning & Land Use	This data set consists of 6 classes of zoning features: zoning districts, special purpose districts, special purpose district subdistricts, limited height districts, commercial overlay districts, and zoning map amendments.
Population by Community Districts	N/A	Public Safety & Social Context	Aggregated population by community districts for the years 1970, 1980, 1990, 2000 and 2010.
Population by Neighborhood Tabulation Area	NTA	Public Safety & Social Context, Environmental & Health Risks	Aggregations of census tracts that are subsets of New York City's 55 Public Use Microdata Areas (PUMAs) of change in population for 2000 and 2010

Crime	Precinct	Public Safety & Social Context, Transportation & Traffic	Statistical breakdown by citywide, borough, and precinct.
Street Construction Permits	Lat/Lon, Street	Construction & Permitting, Transportation & Traffic	Over 150 different types of sidewalk and roadway construction permits to utilities, contractors, government agencies and homeowners.
MTA Subway and Train Lines	Lat/Lon	Transportation & Traffic Comparative Site Queries	A dataset listing all NYC subway and Staten Island Railway stations.
City Owned and Leased Property	BBL	Zoning & Land Use, Construction & Permitting	Identifies city-owned and leased properties and their usage.
Sewer System Data	Point / Polygon	Environmental & Health Risks Comparative Site Queries	This MS4 Map represents the known MS4 outfalls and drainage areas as of August 1, 2020 and provides additional data relevant to the MS4 Permit.
Clean Air Tracking System (CATS)	BBL	Environmental & Health Risks	An online platform managed by the NYC Department of Environmental Protection that facilitates end-to-end processing of air quality.
Water and Sewer Permits	BBL	Construction & Permitting Comparative Site Queries	Contain information about the different types of applications approved and permits issued on a regular basis.
DOB NOW: Build – Job Application Findings	BBL	Construction & Permitting Comparative Site Queries	Records building job filings and applications, with the exception of electrical, elevator, and LAA
Citywide Catch Basins	Lon/Lat	Environmental & Health Risks	The Citywide Catch Basins dataset maps over 150,000 catch basins that collect stormwater and direct it into NYC's sewer network or nearby waterbodies, supporting drainage and flood management.
Parks Monuments	Borough	Zoning & Land Use	This data is a table of monuments that is maintained by NYC Parks.
Citywide Hydrants	Lat/Lon	Public Safety & Social Context	Lists all fire hydrant locations across NYC.

		Comparative Site Queries	
Street Pavement Rating	Borough	Transportation & Traffic	The Agency performs ongoing assessment of New York City streets. Ratings are based on a scale from 1 to 10

### A3: Geounits

Geounit Type	Description	Example Source Dataset
BBL	Parcel-level property identifier	PLUTO
BIN (Building Identification Number)	Building-level identifier	Geoclient
Latitude/Longitude (EPSG:4326)	Coordinate-based	MTA Subway
Precinct	NYPD / enforcement area	Crime Data
NTA (Neighborhood Tabulation Area)	Demographic aggregation	Population by Neighborhood Tabulation Area

### A4: Risk Category

<b>Environmental &amp; Health Risks</b> – asbestos filings, flood zones, noise, and air quality complaints	<b>Zoning &amp; Land Use</b> – zoning districts, landmarks, historic designations, land use restrictions	<b>Construction &amp; Permitting</b> – active DOB building permits, street construction, asbestos filings
<b>Transportation &amp; Traffic</b> – vehicle collisions, traffic volumes, pedestrian injury patterns	<b>Public Safety &amp; Social Context</b> – NYPD complaint types, crime density, demographic context	<b>Comparative Site Queries</b> – multi-location comparisons across any of the above categories

### A5: Midterm Report Evaluation Set

Bucket 1: Exact Matches

Purpose: To test the baseline functionality (Parser, Adapter, Retriever) on clean, unambiguous queries with known data.

1. Is there asbestos risk at 1633 Broadway, New York, NY 10019?
2. What is the asbestos risk at 405 Lexington Avenue?

3. Tell me about asbestos at 111 8th Avenue, New York, NY.
4. asbestos records for 1 World Trade Center
5. Show me the risk profile for 550 Madison Avenue.

#### Bucket 2: Spatial Proximity & Imprecise Queries

Purpose: To test the "Geoscope" and "Record Match Accuracy" on ambiguous spatial language.

6. Is there asbestos risk around 670 Columbus Ave, New York, NY 10025?
7. Show me asbestos reports near 80 Centre Street.
8. What is the risk on the same block as 500 West 120th Street?

#### Bucket 3: Typos & Dirty Data

Purpose: To test the Parsed Address Accuracy and the robustness of the prompt interpreter.

9. asbestos risk for 1633 Broadwy
10. 405 lexington ave
11. 111 8th av
12. 23-01 45th road, queens

#### Bucket 4: Zero-Result & Known Misses

Purpose: To test "Summary Faithfulness" and "Conversational Quality" when no data is found.

13. asbestos at 1 Sheep Meadow, Central Park
14. What is the asbestos risk at 1000 5th Avenue?
15. asbestos records for 123 Main Street, New York, NY

#### Bucket 5: Complex & Comparative Queries

Purpose: To test the identified MVP limitations and the system's ability to handle multi-intent queries.

16. Compare asbestos risk between 405 Lexington Avenue and 1633 Broadway.
17. Which is older, 1633 Broadway or the Chrysler Building?
18. Show me commercial buildings in Midtown with 'A' energy grades.

#### Bucket 6: Guardrail & Out-of-Domain Tests

Purpose: To validate the Data Scoping and ensure the bot rejects out-of-domain questions.

19. What is the weather in Morningside Heights today?
20. Who is the mayor of New York City?

## 9. Acknowledgements

In accordance with course policy, Generative AI tools, including ChatGPT and Cursor, were used to support the writing of this report. These tools helped refine phrasing, enhance clarity, and condense content. All analyses, interpretations, and conclusions were independently developed and validated by our Capstone student team.

We also want to express our gratitude to our mentors from **Town+Gown: NYC @ DDC** for their invaluable guidance and feedback throughout this project.

## 10. References

- [1] X. Zhu, R. Y. M. Li, M. J. C. Crabbe, and K. Sukpascharoen, “Can a chatbot enhance hazard awareness in the construction industry?,” *Frontiers in Public Health*, vol. 10, Nov. 2022. doi: [10.3389/fpubh.2022.993700](https://doi.org/10.3389/fpubh.2022.993700)
- [2] A. Polo-Rodríguez, L. Fiorini, E. Rovini, F. Cavallo, and J. Medina-Quero, “Enhancing Smart Environments with Context-Aware Chatbots using Large Language Models,” *arXiv preprint arXiv:2502.14469*, Feb. 2025. [Online]. Available: <https://arxiv.org/abs/2502.14469>
- [3] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning, “Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback,” *arXiv preprint arXiv:2305.14975*, May 2023. Available: <https://arxiv.org/abs/2305.14975>