GEOCODING CAPITAL PROJECTS: LLM PIPELINE

NYC Department of City Planning | Data Engineering

Harris Wang, Data Fellow | *Emory University, Data Science + Informatics*

Keywords:

Data Pipeline, Natural Language Processing, Generative AI

Summary:

To support the geocoding of capital planning projects from freeform descriptions in the NYC Capital Planning Database (CPDB), Harris collaborated with the DCP Data Engineering team to develop a proof-of-concept (POC) Large Language Model (LLM) transformation pipeline. The pipeline leverages the serverless model provider **Cerebras AI** and generative AI workflow frameworks such as **LangChain**, implemented in **Python** scripts, to deliver an end-to-end process from text to geographic coordinates. Designed with modular components, the POC serves as a foundational skeleton for future development and includes a comprehensive framework for monitoring and evaluating the workflow.

NYC Capital Planning Database:

LLM Pipeline to Geocode Capital Planning Projects

Harris Wang, DCP Data Engineering Team

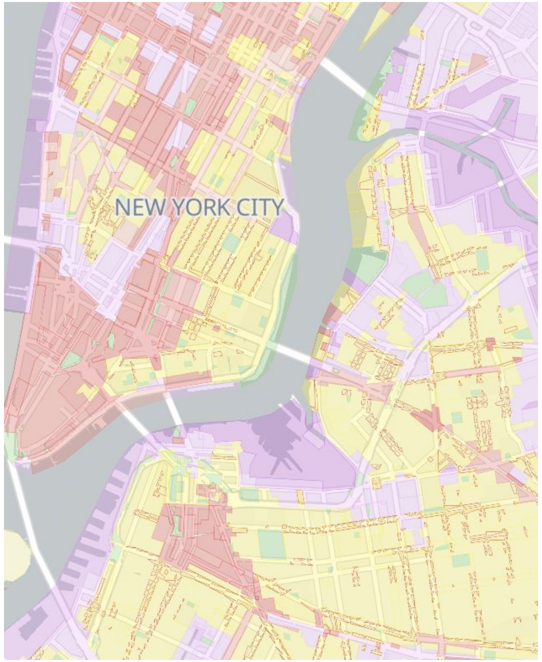
NYC PLANNING

Agenda

- Overview
- Project Scope
- Implementation
- Project Results

Overview

Department of City Planning



The Department of City Planning (DCP) is NYC's primary land use agency

- Responsible for planning construction, growth, and development of NYC
- One of our strategic objectives: Supply data to a broad range of planning functions & stakeholders
- Long history of producing geographic data

NYC PLANNING

Data Engineering Team

Product

Create and release **high quality public datasets** about NYC

Operation

Build highly **transparent** and **automated** data pipelines using **open source technologies**

Ecosystem

Offer more than just data, but also comprehensive **documentation** and **metadata**

Community

Bring people together, across teams and agencies, to share data and to learn from each other

The Data Source

The Capital Projects Database (CPDB) lists current and planned NYC capital projects from the Capital Commitment Plan. A capital project is any public improvement costing at least $50,000 and expected to last five years or more (three years for IT projects).

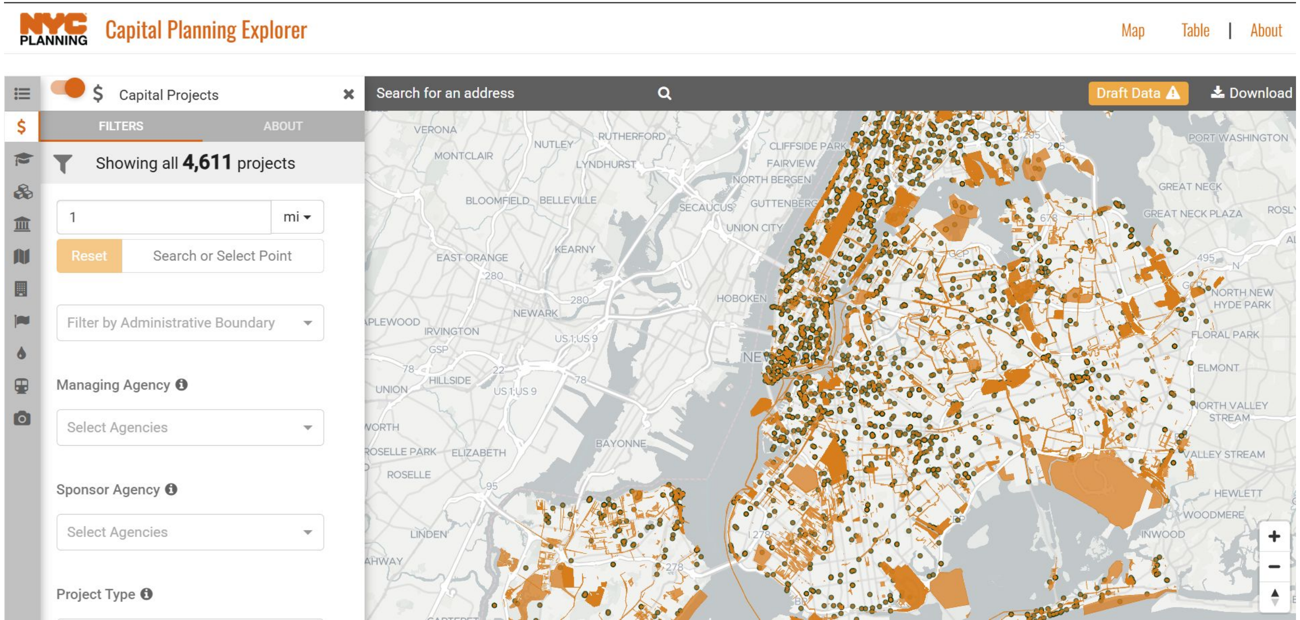
FMS ID ↓	Description	Man. Agency	Spon. Agency	Project Type
035L103RENO	RESEARCH LIBRARIES - Renovations	NYRL	NYRL	New York Research Library
035L19TECHUP	Technology Upgrades - Research Libraries	NYRL	NYRL	New York Research Library
035L20RTECH	Research Libraries - Technology Upgrades	NYRL	NYRL	New York Research Library
035L21EXENNS	NYPL Research Libraries - FY 2022 Executive Plan New Needs	NYRL	NYRL	New York Research Library
035L21FREEZE	NYPL Research Libraries - Blast Freezer	NYRL	NYRL	New York Research Library
035L21JANNNS	NYPL Research Libraries - FY 2022 January Plan New Needs	NYRL	NYRL	New York Research Library
035L21LPAADA	LPA - ADA Lift Replacement	NYRL	NYRL	New York Research Library

No Coordinates Column!!!

NYC PLANNING

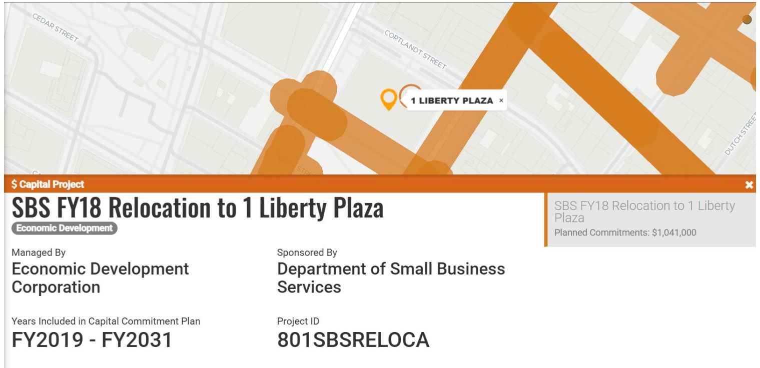
Geocoding the Records

As projects are added and old ones closed, existing geocoding efforts (manual, regex) falls behind: roughly 65 % of projects—about 8,110 out of 12,700—now lack location data in the latest dataset.



Project Goal

SBS FY18 Relocation to 1 Liberty Plaza



Increasing the number of geocoded projects with LLM

NYC PLANNING

What is LLM

Large Language Model (LLM) is a type of artificial intelligence that is trained on massive amounts of text data to generate human-like text and perform various language-related tasks.

deepseek

Claude

Implementation

Project Requirements

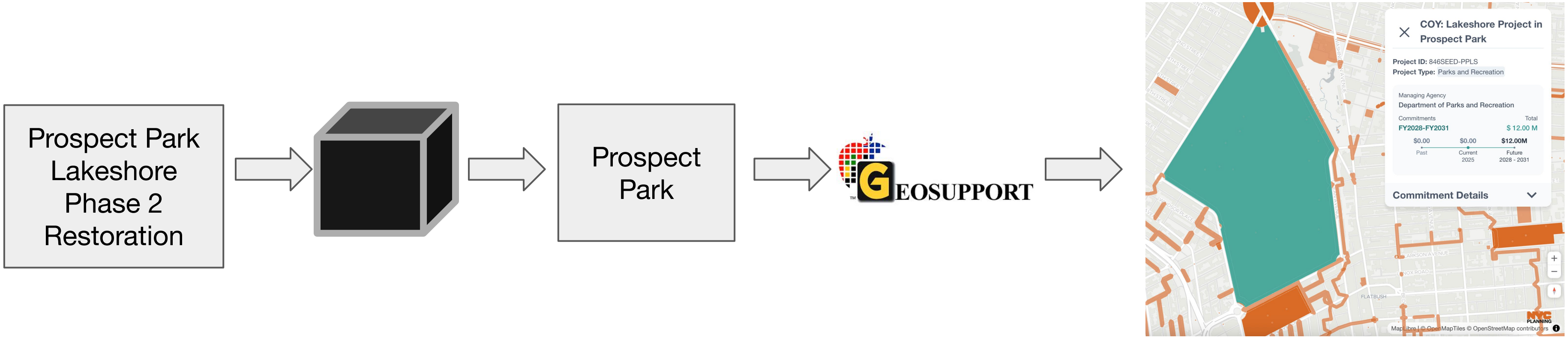
- Light weight
  - Low latency
  - Low computational resources
- Free and open-source
- Accurate
- Secure

LangChain

cerebras

NYC PLANNING

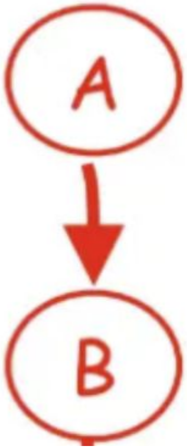
The Pipeline



LLM Workflow: Extract Address

Prospect Park Lakeshore Phase 2 Restoration → “Prospect Park”

Approach 1: Brainstormer + Address Matcher (large volume)



"Dear LLM, plz list clues about location in the project description. Tnx"

"Pinpoint a specific location from the clues. Tnx"

Approach 1:Brainstormer + Address Matcher: Results

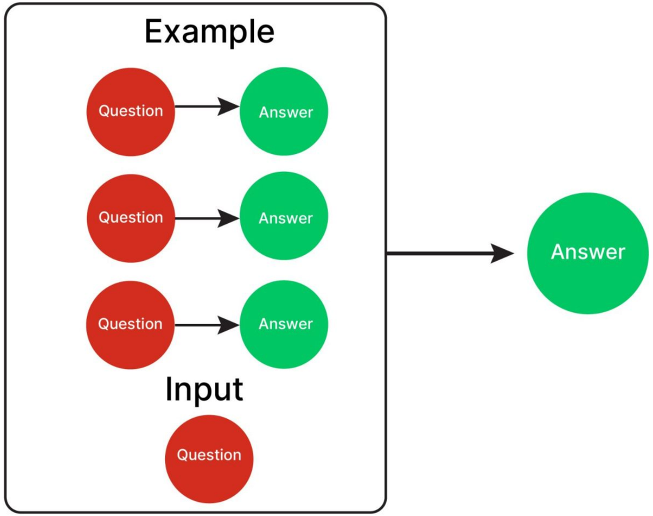
- Ran LLM Pipeline on a geocoded sample of ~300 records
- 40% correct matches
- Out of geocoded records, 60% were incorrectly geocoded (false positives) ← Very HIGH!

```
String Comparison Results: {'total_comparisons': 311, 'exact_matches': 87, 'no_matches': 224, 'false_positives': 151, 'exa
ct_match_rate': 0.2797427652733119}
Output saved to: data\evaluator_string_comparison_geocoded_bbl_output.csv

Geometric Comparison Results: {'total_points': 311, 'valid_comparisons': 238, 'min_distance': 0.0, 'max_distance': 35020.0
2, 'median_distance': 703.28}
```

NYC PLANNING

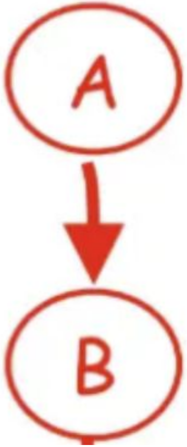
Prompt Engineering: Few Shot



- **Good:** More correctly geocoded records: 36% boost
- **Bad:** Still a lot of false positives: only about 21% decrease

NYC PLANNING

Approach 2: Address Extractor + Verifier (high accuracy)



“Dear LLM, plz extract addresses from the description. Tnx”

“Plz, verify it’s a legit address. Tnx”

Approach 2:Address Extractor + Verifier: Results

- 5% correct matches ← Very low!
- Out of geocoded records, 5% were incorrectly geocoded (false positives) ← Good!

```
String Comparison Results: {'total_comparisons': 311, 'exact_matches': 20, 'no_matches': 291, 'false_posi
tives': 1, 'exact_match_rate': 0.0643086816720257 2}
Output saved to: data\evaluator_string_comparison_geocoded_bbl_output.csv

Geometric Comparison Results: {'total_points': 311, 'valid_comparisons': 21, 'min_distance': 0.49, 'max_d
istance': 29905.64, 'median_distance': 2.92}
```

NYC PLANNING

Summary and Discussion

Noteworthy Results

- Established an LLM framework in a DE setting
  - Modular
  - Efficient
  - Validation Framework
- Assessed different LLM approaches
  - Large volume vs High accuracy

Challenges

Domain knowledge requirement (low volume):

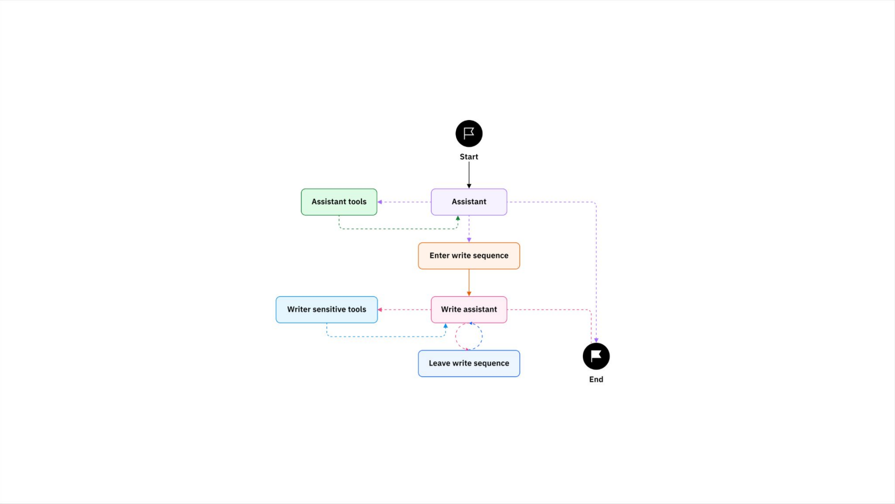
- Most agencies have particular location references rather than actual locations
    - Example. FDNY,KITCHENT RENOVATION - EC316

Hallucination (low accuracy):

- Hallucates frequently when prompted to associate an agency to the an address or pinpoint a landmark

NYC PLANNING

Future Directions



Langraph Agentic Workflow:
Enables more complex reasoning

Web Access:
More contextual cues to assist with predictions

NYC PLANNING

Special Thanks!!!

Alex Richey
Data Engineer

Damon McCullough
Data Engineering Team Lead

Finn van Krieken
Principal Data Engineer

Sasha Filippova
Data Engineer

Project Supervisor

NYC PLANNING

Questions?

NYC PLANNING

Original Approach: Brainstormer + Address Matcher

```
BRAINSTORMER = (
    "You are **Agent 1 (Location Brainstormer)**.\n"
    "\n"
    "Task ▸ From the project description below, list every text span that can "
    "pinpoint a location in NYC, **limited to:**\n"
    "  • Complete street addresses → must have a number *and* a street name "
    "    (e.g. “123 Main St”).\n"
    "  • Named facilities, campuses, parks, bridges, or other unique landmarks "
    "    (“Prospect Park Lakeside Center”, “Brooklyn Bridge”).\n"
    "\n"
    "⚠️ Do **not** include neighborhoods, BBLs, school codes, agency names, "
    "or generic phrases like “various locations”.\n"
    "\n"
    "Output format ▸ Write a **single paragraph** with exactly two labelled "
    "segments in this order (use ‘none’ if empty):\n"
    "    Addresses: <comma-separated addresses>.\n"
    "    Facilities: <comma-separated facilities / landmarks>.\n"
    "End each segment with a period.  Example:\n"
    "    Addresses: 123 Main St, 125 Main St. Facilities: none.\n"
    "\n"
    "Description: {description}\n\n"
    "Clue Paragraph:"
)
```

```
MATCHER = (
    "You are **Agent 2 (Address Matcher)**.\n"
    "\n"
    "From the clue-paragraph below, return **one** location using this order:\n"
    "  1. First address listed after “Addresses:”.\n"
    "  2. If no addresses, the first entry after “Facilities:”.\n"
    "\n"
    "Formatting ▸\n"
    "  • Street address → output exactly “<number> <street>”.\n"
    "  • Facility / landmark → output the name verbatim.\n"
    "  • If both segments are ‘none’, output no results.\n"
    "\n"
    "Think silently, then reply with the single chosen string—no lists, notes, "
    "or commentary.\n"
    "\n"
    "Clue Paragraph: {clues}\n\n"
    "Match:"
)
```

NYC PLANNING

Prompt Engineering: Few Shot

```
examples = [
    {"description": "78 CATHERINE STREET - BATHROOM UPGRADE", "address": "78 CATHERINE STREET"},
    {"description": "101ST PCT HANDICAP RAMP", "address": "no address"},
    {"description": "OCME - Fire Alarm System Upgrade at 520 First Avenue", "address": "520 First Avenue"},
    {"description": "1 CENTRE ST.- OFFICES FROM 49-51 CHAMBERS ST. - 21C3", "address": "1 CENTRE ST"},
    {"description": "RENOVATION OF OLD POLICE ACADEMY/13TH PCT", "address": "no address"},
```

NYC PLANNING

Context Matters

ChatGPT 4o

address of FDNY,KITCHENT RENOVATION - EC316

It looks like “FDNY, KITCHENT RENOVATION – EC316” refers to a specific entry—List No. 316—in the FDNY’s Approved Commercial Cooking Companies List, rather than a building or project location. That entry corresponds to:

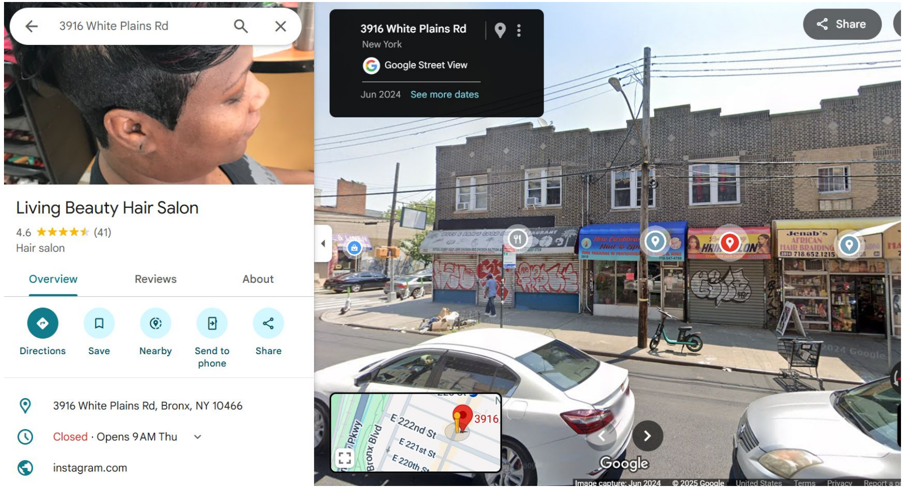
Company: On Demand Fire Protection Inc
Address: 3916 White Plains Road, Bronx, NY 10466 nyc.gov

So, if your question is asking for the address associated with “FDNY, Kitchen Renovation – EC316,” that would be the address above.

If instead you're looking for the FDNY’s physical location or office related to kitchen renovations, or something else entirely—let me know and I’ll happily dig deeper!

Sources

Context Matters



NYC PLANNING

Appendix 1: Evaluator

String Evaluator: Compares matches of 2 string columns (BBL)

- Output in csv with the following columns:
  - Project ID
  - LLM Inferred string column
  - Labelled string column
  - Boolean column indicating matches

Geometric Evaluator: Calculates the distance between 2 geometric columns

- Output in csv with the following columns:
  - Project ID
  - LLM Inferred Geometry
  - Labeled Geometry
  - Distance between the geometry columns (in meters)

NYC PLANNING