

# Evaluating the Feasibility of ChatGPT for Mapping Building Attributes

Qingqing CHEN<sup>a,1</sup>, Linda SEE<sup>b</sup> and Andrew CROOKS<sup>b</sup>

<sup>a</sup>*Department of Geography, University at Buffalo, USA*

<sup>b</sup>*International Institute for Applied Systems Analysis, Austria*

ORCID ID: Qingqing Chen <https://orcid.org/0000-0002-3787-3134>, Linda See

<https://orcid.org/0000-0002-2665-7065>, Andrew Crooks

<https://orcid.org/0000-0002-5034-6654>

**Abstract.** With increasing rates of urbanization, many challenges are emerging regarding urban sustainability such as the energy usage of buildings. Coinciding with this is the growing attention of urban climate models for energy demand estimation and climate adaptation strategies. However, the applicability of these models is constrained by the lack of detailed urban surface information. Therefore, creating comprehensive datasets that capture urban surface information at a granular scale is crucial for responding to our rapidly urbanizing world. Recent advancements in Multimodal Large Language Model (MLLMs) have opened new opportunities in urban studies, offering accessible methods for information extraction. In this chapter we explore the feasibility of ChatGPT to extract building attributes from images. Taking New York City as a case study, we collect building images from Street View Imagery and process them through ChatGPT by posing specific questions to extract building attributes (e.g., height, functions, age). These attributes are then compared with authoritative data. The proposed method helps address the current dearth of fine-grained surface data on urban issues, therefore enhancing the accuracy and utility of urban climate models. Overall, this study demonstrates the practical applications of ChatGPT in geographic knowledge extraction, advancing the understanding of MLLMs in geographic contexts, and more broadly to the discourse on Artificial Intelligence (AI) in urban modeling and climate science.

**Keywords.** Buildings, ChatGPT, Multimodal Large Language Models (MLLMs), Mapillary, Street View Imagery (SVI)

## 1. Introduction

As the spatial resolution of climate models becomes finer due to increases in computing power and storage, there is a need to represent urban areas and local scale urban processes with greater detail than ever before [1]. Although urban areas account for only 3% of the Earth's surface, they have impacts in the form of urban heat islands, increased runoff from impervious surfaces and greater air pollution [2], while collectively accounting for around 70% of global CO<sub>2</sub> emissions [3]. Moreover, with 55% of the world's population

---

<sup>1</sup>Corresponding Author: Qingqing Chen, [qchen47@buffalo.edu](mailto:qchen47@buffalo.edu).

currently living in urban areas and this number expected to increase to 68% by 2050 [4], understanding and managing these urban impacts becomes even more critical.

On the city scale, urban climate models are now being used for different applications such as determining the impact of climate change in the future and modeling the effects of climate change adaptation and mitigation measures [5, 6]. However, to represent urban areas, much more information about cities is needed that captures microclimatic conditions, including their form and function. By form we mean the shape and configuration of the city, which is captured through building attributes such as height and building materials as well as the spatial layout of the buildings and other infrastructure. In contrast, function refers to the activities that take place in a city, which can be captured through building function such as residential or commercial use [7].

To infer some of the form parameters, the Local Climate Zone (LCZ) classification was developed. Using Landsat imagery, cities are classified into 10 urban types and 7 natural types, which are defined by the height of the buildings and vegetation as well as the density of the buildings [8]. This approach has been applied globally to aid the parameterization of urban areas in climate models [9]. However, LCZs provide only some of the information needed. Another form parameter of importance is the age of the building, as this has been shown to be a key factor in modeling the energy consumption of buildings [10]. Using various urban morphological parameters as inputs, research into predicting building age with machine learning has yielded accuracies of 77% for buildings in the city of Nottingham, UK [11] and mean absolute errors of between 15 and 20 years for buildings in the Netherlands and Spain [12].

In terms of function, knowing whether a building is residential, commercial or mixed use can also provide information on energy usage and carbon emissions, yet such information is not readily available at the city scale. For example, the Points of Interest (POIs) in OpenStreetMap provide one source of building function [see, e.g., 13] although there are issues with both completeness and consistency in tagging between cities, while another study has used POIs in combination with taxi trajectories and WeChat user locations to infer building function in Guangzhou, China [14]. However, a more comprehensive and harmonized data set on building function is not available.

Street level imagery such as Google Street View [15] and Mapillary [16] provide a new, valuable source of building images from which form and function can be extracted. A number of studies have appeared that have used computer vision and deep learning approaches including convolutional neural networks (CNNs) to determine building function, height and age. For example, [17] used pre-trained CNNs and Google Street View imagery to predict different building types including residential and commercial for buildings in Calgary, Boston and Toronto. The F1 scores varied between 0.65 and 0.82 for residential while scores were lower for commercial. Different approaches have been used to estimate building height but in a comparison using the same data set, [18] showed that a pre-trained CNN using Google Street View imagery for Toronto outperformed other approaches including those based on geometry, with a mean absolute error of 1.2 m. Similarly, there have been studies that have attempted to predict building age. For example, [19] used a combination of a CNN to extract features from street level imagery and a support vector machine to predict the year of construction. The mean absolute error varied between 10 to 12 years depending on the pre-trained CNN used. Finally, [20] used a CNN to predict building age from street level imagery from Amsterdam, re-

sulting in an accuracy of 81%. However, all of these studies use some form of pre-trained CNN and/or other machine learning approaches.

There are two new innovations that could potentially be used to help fill the information gap on building form and function. The first is new open data sets that have emerged on building footprints, heights and building age while the second is advances in generative Artificial Intelligence (AI) in the form of Multimodal Large Language Models (MLLMs) that have been trained on vast amounts of information [21]. Unlike traditional LLMs that primarily process and generate text, MLLMs are capable of understanding and generating information across different modes, including text and images, thus can be used to describe what they see in photographs [see, e.g., 22]. Such an approach has not yet been used to determine whether information on building height, age and function can be extracted from photographs using MLLMs. Hence, the aim of this chapter is to investigate the feasibility of this approach. Demonstrated using a sample of buildings across the city of New York from Mapillary images, which have been matched to building footprints using a spatial-analytical approach and then classified using ChatGPT, we highlight the advantages and limitations of such an approach for creating wall-to-wall data sets on building attributes in the future.

## 2. Methodology

### 2.1. Study Area and Data

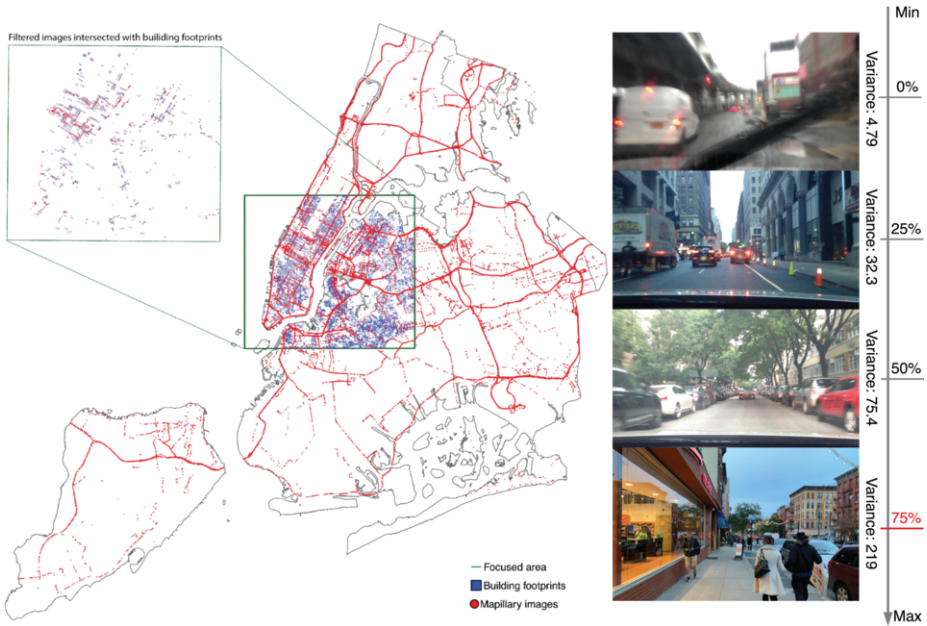
To illustrate the proposed idea, we utilized street view images of New York City (NYC) collected from Mapillary through its online Application Programming Interface (API), which comprises approximately 2.5 million images from 2018. Each image contains four key attributes: a unique identifier, a timestamp indicating the time of capture, geographical coordinates and a compass angle reflecting the direction of capture. Figure 1 shows the spatial distribution of a 2% sample of these images, with a higher density in Manhattan. This led us to focus our analysis on this area (outlined in the bounding box), where we filtered images intersecting with building footprints, yielding a total of 13,708 images. However, a common issue with such Volunteered Street View Imagery (VSVI) is its varying image quality, as not all images are of high resolution. To address this, we calculated the derivative of the images using a Laplacian Operator—a technique commonly used for edge detection—to discern high-resolution images by selecting those within the upper quartile of image variance (Variance  $\geq 219$ ), highlighted in red in Figure 1. This process eliminated around 75% of the images, resulting in a refined subset of high-resolution 3,427 images.

### 2.2. Ground Truth Data Collection

As ready-to-use ground truth data are not available due to the distributed nature of these data, we constructed a bespoke ground truth data set for building height and year of construction by integrating data from multiple sources. Specifically, we performed a spatial join between two key datasets: the *Building Heights*<sup>2</sup> dataset, which provides detailed

---

<sup>2</sup>The *Building Heights* dataset, provided by the New York City Department of Information Technology & Telecommunications (DoITT), offers point-based height and construction year information for buildings across the city. This dataset is part of the *NYC Open Data* initiative.



**Figure 1.** The spatial distribution of Mapillary images within the study area, shown on the left, and the distribution of images by variance showing increasing image quality on the right.

point-based height and year of construction information across New York City, and the *USBuildingFootprints*<sup>3</sup> dataset, which contains computer generated building footprint polygons for the United States, derived from satellite imagery. This process allowed us to assign both building height and year of construction to each building footprint polygon. However, in some cases, the *USBuildingFootprints* dataset has combined multiple adjacent buildings into one polygon, which limits the ability to assess individual building characteristics. Therefore, to ensure that each building has its own distinct polygon, we improved the dataset using the latest release of *Overture*<sup>4</sup> building data from Overture Maps to refine the building geometries. Overture Maps integrates multiple data sources, such as OpenStreetMap and Google’s Open Buildings [23], thereby providing more accurate and distinct building outlines than the solely machine-generated Microsoft footprints, which often fails in detecting small buildings smaller than  $100m^2$  [24].

Additionally, to generate ground truth labels for building types, we conducted a manual labeling process via an online questionnaire. The survey consisted of 92 images that were randomly sampled from the refined subset of high-resolution images, where participants were tasked with classifying the leftmost and rightmost buildings in each image into one of three categories: ‘Residential’, ‘Commercial’, or ‘Mixed-Use’. The ‘leftmost’ building refers to the building closest to the left edge of the image, while the ‘rightmost’ building is the building nearest to the right edge. Two domain experts completed the la-

<sup>3</sup>The *USBuildingFootprints* dataset is provided by Microsoft and consists of over 160 million building footprints across the United States.

<sup>4</sup>*Overture Maps* provides free and open map data, harmonized to a common schema with themes such as buildings, addresses and transportation. It includes over 3.7 billion features sourced from OpenStreetMap, Meta, Microsoft, Esri and others.

being process independently. In cases where discrepancies arose between their labels, a third expert facilitated a discussion to reach a consensus, ensuring a final classification that all three experts agreed upon. These labels, together with the building height and year of construction of the buildings, formed the final ground truth dataset that was used to evaluate the same building information extracted from the images by ChatGPT.

### 2.3. Implementation

Once the images were collected and the ground truth dataset was established, we then implemented a workflow to extract building attributes using ChatGPT model, specifically, the ‘gpt-4-turbo’ model, through the OpenAI API in July and August 2024. The images were processed in batches to enhance the efficiency, with ChatGPT providing building type, height, and construction period for buildings in the images. To ensure clear guidance to ChatGPT, we refined the prompts through multiple iterations. For example, we engaged with ChatGPT to define terms like ‘leftmost’ accurately by conducting several test runs, asking it to describe the leftmost building in each image until it consistently identified the correct one. This process resulted in a structured prompt, instructing ChatGPT to analyze both the leftmost and rightmost buildings in each image and return a JSON formatted response with a brief description of the building of interest, estimated height in feet, primary use, and construction period. Details of the final parameters used in the model are presented in Table 1 while Figure 2 gives an overview of the research workflow.

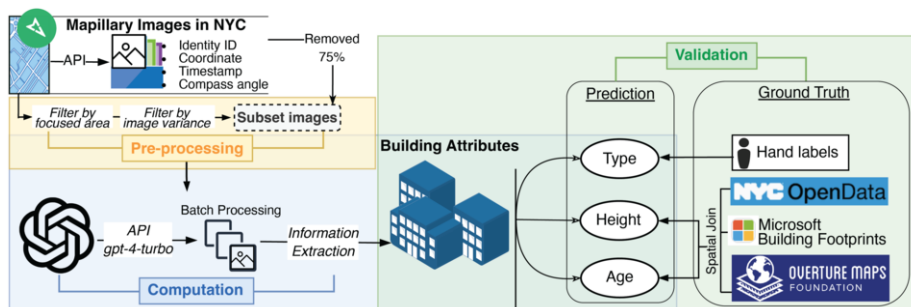


Figure 2. An overview of the research workflow.

## 3. Results and Discussion

We evaluated ChatGPT’s performance in extracting building attributes using accuracy and F1 score metrics. The accuracy measures the overall correctness of the predictions, while the F1 score balances precision and recall, offering a detailed assessment of the model’s performance in identifying building characteristics. In what follows, we will detail ChatGPT’s feasibility in classifying building types, determining the period of construction and identifying building heights, along with a discussion of its strengths and limitations.

Parameter	Description
Model	gpt-4-turbo
Temperature	0.2 (Controls the randomness of the response, with lower values making the output more focused and deterministic.)
Max Tokens	350 (Specifies the maximum length of the response)
Position	leftmost OR rightmost
Position direction	left OR right
System message	You are an AI trained to analyze building images. Each image may contain multiple buildings. When referring to buildings, {position} means the building located furthest to the {position direction} in the image. Provide a short description of the building, including its key visual characteristics, as well as information on building height, primary use, and construction period.
User message	For the {position} building, which is the building located furthest to the {position direction} in the image (i.e., the building that appears closest to the {position direction} edge of the image), please provide the following details: 1. A short description of the building, including its key visual characteristics; 2. The approximate height of this building in feet; 3. The primary use of the building: 1 residential, 2 for commercial, 3 for mixed purposes, just return a number between 1 to 3; 4. The estimated period when this building was constructed, as a range: 1 Victorian (1840-1900), 2 Pre-WWII (1900-1946), 3 Post-WWII (1946-1970), 4 Postmodernist (1970-2000), 5 Contemporary (2000-2024). Please return the output in a JSON object contain the following information: <pre>{   description: string,   height_feet: number,   building_use: number,   construction_period: number }</pre>

**Table 1.** Detailed description of the parameters used in processing the images with the OpenAI API.

### 3.1. Building Type

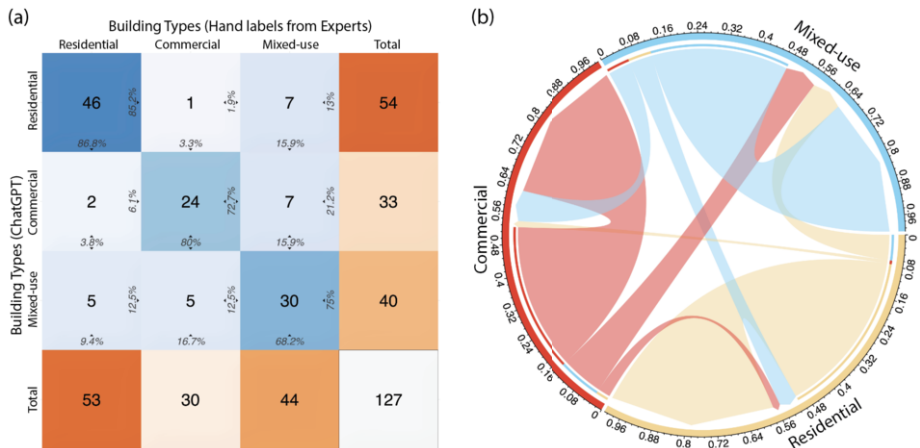
Table 2 presents a comparison between the manually-labeled building types and the ChatGPT classifications across the three different building types. The results show that ChatGPT demonstrates a strong performance in classifying these building types, with an overall accuracy of 0.79 and an F1 score of 0.86. The model performs the best at identifying residential buildings, with a high F1 score (0.86), while a slightly lower performance was achieved for commercial ( $F1 = 0.76$ ) and mixed-use buildings ( $F1 = 0.68$ ).

Building Type	N	Sensitivity	Specificity	Precision	Recall	F1
Residential	54	0.87	0.89	0.85	0.87	0.86
Commercial	33	0.80	0.91	0.73	0.80	0.76
Mixed-use	40	0.68	0.88	0.75	0.68	0.71
		Overall accuracy: 0.79		Overall F1: 0.86		

**Table 2.** Statistical summary of building types classified by ChatGPT compared to the manually-labeled building types from the experts.

To investigate those cases where ChatGPT misclassified the building types, we further break down the classification patterns. The confusion matrix (see Fig. 3(a)) high-

lights that while 87% of residential buildings were correctly identified, 9% were mislabeled as mixed-use and 4% as commercial. Similarly, for commercial buildings, 24 out of 30 were correctly classified, but 5 were misidentified as mixed-use, indicating a certain overlap between these two categories. Mixed-use buildings showed the highest misclassification rate, with 32% (14 out of 44) incorrectly labeled as either residential or commercial. The chord diagram in Figure 3(b) visually emphasizes these trends, showing a high degree of agreement for residential buildings and greater confusion between the mixed-use and commercial categories.



**Figure 3.** Comparison of building type classifications. (a) A confusion matrix detailing the distribution of ChatGPT's classifications against the hand labels from experts; (b) A chord diagram illustrating the proportion of classifications for each building types as labeled by experts compared to ChatGPT's classifications.

An in-depth review of the building description provided by ChatGPT revealed that the model classifies buildings based on visual cues and contextual information, rather than official zoning or governmental standards. For example, residential buildings (i.e., houses, apartments, and residential complexes) are identified by features such as windows, balconies, and private entrances commonly seen in residential neighborhoods. In contrast, commercial buildings are recognized by larger windows (or lack of windows), signage and visible business activities. Mixed-use buildings show a combination of residential and commercial characteristics, often featuring retail or office spaces at street level with residential units above. Figure 4 provides examples of buildings where ChatGPT's classification aligns with the expert labels.

To further explore the underlying reasons for these misclassifications, we examined instances of where ChatGPT's predictions diverged from the ground truth and then identified three main challenges encountered by the algorithm, which we refer to here as 'Building Boundary Confusion', 'Dominance of Visual Features' and 'Visual Obstruction-induced Assumptions'. Figure 5 provides examples of buildings corresponding to these three problem areas. The 'Building Boundary Confusion' challenge arises when ChatGPT merges visual cues from closely positioned buildings, misinterpreting them as a single entity. For example, in Figure 5 (Image ID: 479109580026052 & 953148148851651), the model combines commercial signage from one building and residential characteristics from another, resulting in a mixed-use classification. The 'Domi-



**Figure 4.** Examples of buildings with matched building type labels.

nance of Visual Features’ issue occurs, for example, when dominant residential features (e.g., fire escapes) overshadow subtler commercial indicators (e.g., storefronts), leading to misclassifications (see Image ID: *139601918186023* in Figure 5).

Finally, the ‘Visual Obstruction-induced Assumptions’ relates to a common quality issue of street view image–obstruction [e.g., 25], which occurs when a portion of the image has been blocked by certain objects, such as vehicles, signboards or people, forcing the model to make assumptions based on incomplete visual cues. For example, in Figure 5 (Image ID: *1157537391430203* & *505936270580700*), ChatGPT misclassified the buildings as mixed-use when they were actually labeled as residential due to obstructed ground-level views by parked cars. In these cases, ChatGPT may infer the building type based on assumptions related to the visible facade or the upper stories of the building, leading to an incorrect categorization.

### 3.2. Construction Period of the Building

In the extraction of building age, instead of specifying the exact construction years, we opted to classify buildings into five historical periods: ‘Victorian (1840-1900)’, ‘Pre-WWII (1901-1946)’, ‘Post-WWII (1947-1970)’, ‘Postmodernist (1971-2000)’ and ‘Contemporary (2001-2024)’. Each category represents distinct architectural characteristics that are associated with these historical epochs. Specifically, the Victorian period, driven by the Industrial Revolution and the growth of the middle class, produced diverse architectural styles, from functional terraced homes to more elaborate residences, often distinguished by steeply pitched roofs, ornate gables, and asymmetrical facades. In contrast,







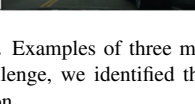
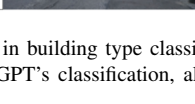
Challenges	Examples	
Building Boundary Confusion	 <p>Image ID: 479109580026052 Ground Truth (L1) Commercial ChatGPT Prediction (L1) Mixed-use The building hosts a commercial business on the ground floor, indicated by the visible pawn shop signage, and likely contains residential units above, as suggested by the multiple stories and fire escapes. The combination of uses is common in urban areas.</p>	 <p>Image ID: 953148148851651 ChatGPT Prediction (L1) Commercial Ground Truth (L1) Mixed-use The building is used for mixed purposes. The presence of commercial signage and a storefront on the ground floor indicates commercial use, while the upper floors, with large windows typical of residential units, suggest residential use.</p>
	 <p>Image ID: 139601918186023 Ground Truth (R1) Mixed-use ChatGPT Prediction (R1) Residential The building's external fire escapes and dense window distribution suggest it is primarily used for residential purposes.</p>	 <p>Image ID: 1461128657581037 Ground Truth (L1) Mixed-use ChatGPT Prediction (L1) Residential The building's narrow design and presence of multiple fire escapes suggest it is primarily used for residential purposes, typical of such structures in densely populated urban areas.</p>
Dominance of Visual Features	 <p>Image ID: 1157537391430203 Ground Truth (L1) Residential ChatGPT Prediction (L1) Mixed-use The leftmost building is a two-story structure with a brick facade and visible signage. It features a flat roof and several commercial signs, indicating mixed-use with heavy shops on the ground floor.</p>	 <p>Image ID: 505936270580700 Ground Truth (L1) Residential ChatGPT Prediction (L1) Mixed The building's multiple stories with residential-style windows suggest mixed use, possibly containing both residential units and small commercial or office spaces on the lower floors.</p>
Visual Obstruction-Induced Assumptions		

Figure 5. Examples of three major challenges identified in building type classification with ChatGPT. For each challenge, we identified the ground truth and ChatGPT’s classification, along with returned building description.

Pre-WWII architecture is marked by its brick or stone facades, decorative cornices and intricate detailing around windows and doors, reflecting a focus on craftsmanship and quality materials. Post-WWII buildings, constructed in response to economic constraints and rapid urban development, shifted towards more functional and minimalist designs. The Postmodernist period introduced eclectic and playful architectural elements, reinterpreting classical elements such as columns and arches in bold and unconventional ways. Contemporary architecture, in contrast, focuses on sustainability and technological innovation, with features like large glass windows, eco-friendly materials and smart building systems.

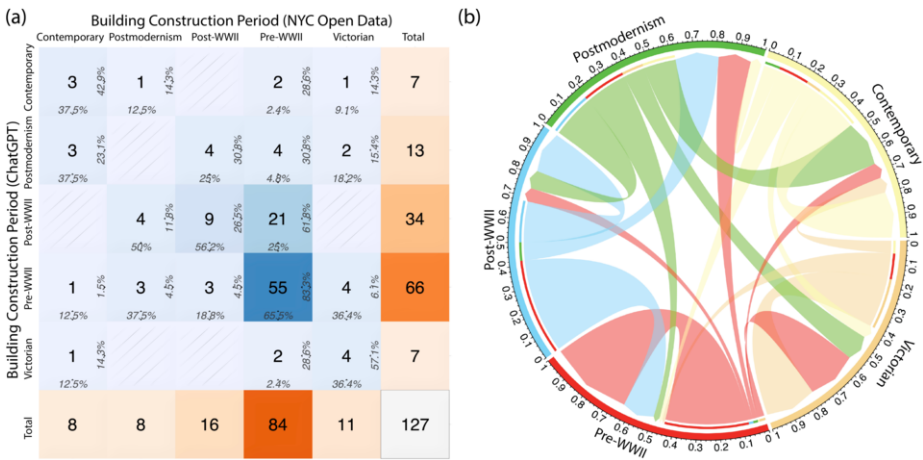
The rationale for adopting these broad categories is that architectural styles evolve gradually over decades rather than changing abruptly. In this sense, buildings from the same period often share similar visual and structural features, making it challenging for ChatGPT to determine precise construction years from visual cues alone. To circumvent this issue, we thus grouped years of construction into defined time periods for more reliable classification.

Table 3 summarizes ChatGPT’s performance in predicting construction periods of buildings. Overall, the model achieved moderate accuracy (0.56) and an F1 score of 0.44. Among the five categories, the ‘Pre-WWII’ category exhibits the best performance ( $F1 = 0.73$ ), followed by the ‘Victorian’ ( $F1 = 0.44$ ) and ‘Contemporary’ ( $F1 = 0.40$ ) periods. However, ChatGPT had difficulty in differentiating between ‘Post-WWII’ and ‘Postmodernist’ buildings, indicating the potential overlap in visual features.

Building Age Period	N	Sensitivity	Specificity	Precision	Recall	F1
Victorian [1840-1900]	11	0.36	0.97	0.57	0.36	0.44
Pre-WWII (1900-1946)	84	0.65	0.74	0.83	0.65	0.73
Post-WWII (1946-1970)	16	0.56	0.77	0.26	0.56	0.36
Postmodernist (1970-2000)	8	0.00	0.89	0.00	0.00	-
Contemporary (2000-2024)	8	0.38	0.97	0.43	0.38	0.40
		Overall accuracy: 0.56		Overall F1: 0.44		

Table 3. Statistical summary of the period of construction identified by ChatGPT.

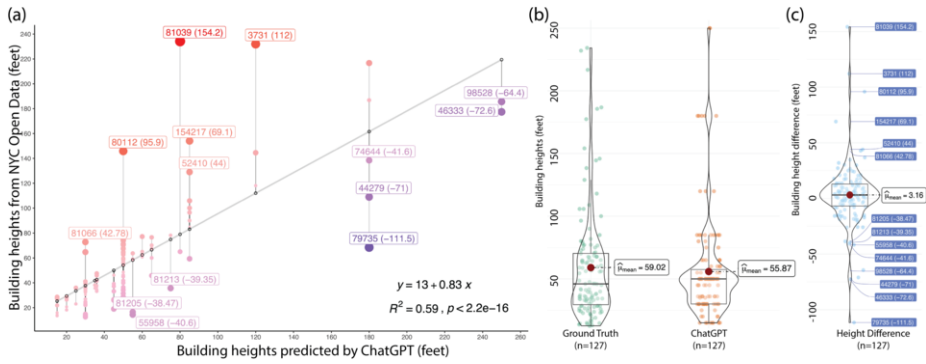
To further investigate age classification limitations, Figure 6 highlights specific cases of misclassification. For example, while ‘Pre-WWII’ buildings showed the highest F1 score, 21 out of 84 buildings (25%) were misclassified as ‘Post-WWII’. This misclassification seems to stem from overlapping architectural features between adjacent periods since buildings from adjacent periods may share certain visual similarities, which can be seen in the confusion matrix (see Fig.6(a)). This could confuse the model, therefore, leading to ambiguity when distinguishing specific period features. Additionally, historical renovations and modifications can obscure original features, further complicating the classification. Similar confusion was observed between ‘Victorian’ and ‘Pre-WWII’ categories. Conversely, the ‘Postmodernist’ category shows sparse agreement, with confusion spreading to all other categories. The chord diagram (Fig.6(a)) visually represents these classification challenges and the patterns of agreement and confusion among the categories.



**Figure 6.** Comparison of the building period of construction from the ground truth data and the classifications from ChatGPT. (a) A confusion matrix which details the distribution of buildings classified within each period by ChatGPT compared to the ground truth data; (b) A chord diagram illustrating the patterns of agreement and confusion among the categories.

### 3.3. Building Height

Turning to building height, Figure 7(a) demonstrates the relationship between the heights predicted by ChatGPT and those recorded in NYC Open Data. The results show a moderate correlation ( $R^2 = 0.59$ ) between the two, indicating that 59% of the variance in the actual building heights can be explained by the predicted heights from ChatGPT. The red labels represent outliers where ChatGPT underestimated the height, whereas purple labels indicate an overestimation. Each label provides the building ID and the height difference for context. Given that building height is continuous data, we opted to use the  $R^2$  metric for evaluating the results, as it is more suitable for assessing the accuracy of predictions in this context than F1 scores, which are typically used for binary classification tasks. Figure 7(b) presents the distribution of the building heights from NYC Open Data and ChatGPT, with the average predicted height (56 feet) closely matching the ground truth (59 feet) and a mean deviation of approximately 3 feet (see Fig. 7(c)).



**Figure 7.** A comparison of building heights from ChatGPT and the NYC Open Data. (a) The correlation of height between the ground truth and ChatGPT; (b) The distribution of ground truth heights and the predicted heights; (c) The difference in the heights.

To better understand the discrepancies in height prediction, we examined outliers and identified two main challenges that affect ChatGPT’s performance, which we refer to here as: ‘*Incomplete Visibility*’ and ‘*Building Boundary Confusion*’ (see Figure 8). In particular, ‘*Incomplete Visibility*’ challenge arises when only part of a building is visible in an image. In these cases, ChatGPT needs to make an estimation based on a limited visible portion of the building, often resulting in an underestimation. This issue is exacerbated by the angle and distance from which the photo is taken, which can distort the perceived scale. The ‘*Building Boundary Confusion*’ issue, also noted earlier in the building type classification, occurs when adjacent buildings appear visually connected, leading ChatGPT to mistakenly interpret them as a single entity.

Challenges	Examples		
Incomplete Visibility			
	Image ID: 1367217583642843	Image ID: 469584784104135	Image ID: 767036650847795
	Ground Truth (R1) 72.78 ChatGPT Prediction (R1) 30	Ground Truth (L1) 154.1 ChatGPT Prediction (L1) 85	Ground Truth (R1) 234.2 ChatGPT Prediction (R1) 80
Building Boundary Confusion			
	Image ID: 293347652339995	Image ID: 479109580026052	Image ID: 953148148851651
	Ground Truth (R1) 68.5 ChatGPT Prediction (R1) 180	Ground Truth (L1) 16.53 ChatGPT Prediction (L1) 55	Ground Truth (R1) 14.4 ChatGPT Prediction (R1) 55

**Figure 8.** Examples of two major challenges identified in building height determination using ChatGPT.

## 4. Conclusions

This chapter demonstrated the potential of ChatGPT for identifying characteristics of urban form and function from street level photographs, taken from Mapillary. The most successful example of using ChatGPT was for determining building function, which is a dataset that does not openly exist on a large scale. Hence there is considerable potential for upscaling this approach and contributing to open datasets like Overture with information on building function. This data could be integrated into other spatial or geographic data, such as zoning information, to analyze energy consumption, property pricing, regulatory compliance, among others. Building on this, the study focused on NYC, and from this perspective, it is essential for future work to apply the proposed method to other cities or regions. In other words, the methodology, informed by the specific nature of the data and academic context, is by definition contingent and might require adjustments when applied to different geographical settings or data types.

Building height and age were also determined with reasonable accuracy, which could be improved through modifying the interactions with ChatGPT and handling the challenges identified. For example, ‘Building Boundary Confusion’ and ‘Visual Obstruction-induced Assumptions’ could be overcome by refining the building selection from the street view image. Using higher quality street view images might overcome these challenges. For example, with Mapillary, we are currently restricted to where the camera is pointing, which is often driving a down street, while commercial street view imagery tends to offer 360 degree panoramas. In addition to the quality of the street view imagery, we can also improve the reliability of the responses by adding extra steps in the prompts. For example, before estimating the building heights, we could ask ChatGPT to determine whether it can see the top of the building. Similarly, for the ‘Visual Obstruction-induced Assumption’ and ‘Incomplete Visibility’, we can ask ChatGPT to disregard the images that have these issues. Moreover, to enrich the robustness of our results, future work could focus on testing different models (e.g., Google, OpenAI, Llama) with differing parameter sizes and initial training data. Another area of research would be to compare these MLLMs with more traditional computer vision models.

However, even with these challenges, ChatGPT and other emerging MLLMs offer new opportunities to not only extract geographic knowledge from imagery, but for parameterizing climate models. For example, by using different prompts, one could envisage the use of MLLMs to inform the LCZ classification. Another key advantage of MLLMs for studying the form and function of cities at scale is their accessibility. Unlike traditional approaches that require extensive expertise in computational analysis, such as computer vision and deep learning [26, 27], MLLMs leverage pre-trained capabilities that simplify the analytical process. This accessibility enables a larger group of researchers to incorporate MLLMs in their analyses. This, of course, needs to be done in a critical and cautious manner to ensure that the robustness of the methodology is adequately considered at the research implications stage.

## References

- [1] Ashish Sharma, Donald J Wuebbles, and Rao Kotamarthi. The need for urban-resolving climate modeling across scales. *AGU Advances*, 2(1):e2020AV000271, 2021.

- [2] A Baklanov, C Sue B Grimmond, D Carlson, D Terblanche, X Tang, V Bouchet, B Lee, G Langendijk, R Kumar Kolli, and A Hovsepyan. From urban meteorology, climate and environment research to integrated city services. *Urban Climate*, 23:330–341, 2018.
- [3] IPCC. Summary for Policymakers. In *Climate change 2022: Mitigation of climate change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khouradajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]*, pages 3–48. Cambridge University Press, Cambridge, UK and New York, NY, USA, 1 edition, 2023.
- [4] United Nations. 68% of the world population projected to live in urban areas by 2050, says UN, 2025. Publisher: United Nations.
- [5] Rafiq Hamdi, Hiroyuki Kusaka, Quang-Van Doan, Peng Cai, Huili He, Geping Luo, Wenhui Kuang, Steven Caluwaerts, François Duchêne, Bert Van Schaeybroek, et al. The state-of-the-art of urban climate change modeling and observations. *Earth Systems and Environment*, 4:631–646, 2020.
- [6] Sandro M Oswald, Brigitta Hollosi, Maja Žuvela-Aloise, Linda See, Stefan Guggenberger, Wolfgang Hafner, Gundula Prokop, Alexander Storch, and Wolfgang Schieder. Using urban climate modelling and improved land use classifications to support climate change adaptation in urban environments: A case study for the city of klagenfurt, austria. *Urban Climate*, 31:100582, 2020.
- [7] Jelena Živković. Urban form and function. *Climate action*, pages 862–871, 2020.
- [8] Benjamin Bechtel, Paul J Alexander, Jürgen Böhner, Jason Ching, Olaf Conrad, Johannes Feddema, Gerald Mills, Linda See, and Iain Stewart. Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, 4(1):199–219, 2015.
- [9] Matthias Demuzere, Jonas Kittner, Alberto Martilli, Gerald Mills, Christian Moede, Iain D Stewart, Jasper Van Vliet, and Benjamin Bechtel. A global map of local climate zones to support earth system modelling and urban scale environmental science. *Earth System Science Data Discussions*, 2022:1–57, 2022.
- [10] Mehmet Aksoezen, Magdalena Daniel, Uta Hassler, and Niklaus Kohler. Building age as an indicator for energy consumption. *Energy and Buildings*, 87:74–86, 2015.
- [11] J.F. Rosser, D.S. Boyd, G. Long, S. Zakhary, Y. Mao, and D. Robinson. Predicting residential building age from map data. *Computers, Environment and Urban Systems*, 73:56–67, 2019.
- [12] Florian Nachtigall, Nikola Milojevic-Dupont, Felix Wagner, and Felix Creutzig. Predicting building age from urban form at large scale. *Computers, Environment and Urban Systems*, 105:102010, 2023.
- [13] CC Fonte, M Minghini, V Antoniou, J Patriarca, and L See. Classification of building function using available sources of VGI. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:209–215, 2018.
- [14] Ning Niu, Xiaoping Liu, He Jin, Xinyue Ye, Yu Liu, Xia Li, Yimin Chen, and Shaoying Li. Integrating multi-source big data to infer building functions. *International Journal of Geographical Information Science*, 31(9):1871–1890, 2017.

- [15] Google. Explore street view and add your own 360 images to Google Maps., 2025.
- [16] Mapillary. Mapillary, 2025.
- [17] Jian Kang, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:44–59, 2018.
- [18] Alexander W Olson and Shoshanna Saxe. Single-image building height estimation using EfficientNet: A simplified, scalable approach. *Findings*, 2024.
- [19] Yan Li, Yiqun Chen, Abbas Rajabifard, Kourosh Khoshelham, and Mitko Aleksandrov. Estimating building age from Google Street View images using deep learning. In Stephan Winter, Amy Griffin, and Monika Sester, editors, *LIPICs, GIScience 2018*, volume 114, pages 40:1–40:7. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.
- [20] Maoran Sun, Fan Zhang, and Fabio Duarte. Automatic building age prediction from street view images. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 102–106, Beijing, China, 2021. IEEE.
- [21] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.
- [22] Andrew Crooks and Qingqing Chen. Exploring the new frontier of information extraction through large language models in urban analytics, 2024.
- [23] Overture Maps. Buildings | Overture Maps Documentation, 2025.
- [24] Mehdi P. Heris, Nathan Leon Foks, Kenneth J. Bagstad, Austin Troy, and Zachary H. Ancona. A rasterized building footprint dataset for the United States. *Scientific Data*, 7(1):207, 2020. Publisher: Nature Publishing Group.
- [25] Yujun Hou and Filip Biljecki. A comprehensive framework for evaluating the quality of street view imagery. *International Journal of Applied Earth Observation and Geoinformation*, 115:103094, 2022.
- [26] Yue Wu, Luke S Blunden, and AbuBakr S Bahaj. City-wide building height determination using light detection and ranging data. *Environment and Planning B*, 46(9):1741–1755, 2019. Publisher: SAGE Publications Ltd STM.
- [27] Yizhen Yan and Bo Huang. Estimation of building height using a single street view image via deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:83–98, 2022.