Ritter *et al.*
Frontiers of Urban and Rural Planning (2025) 3:7
https://doi.org/10.1007/s44243-025-00057-2

Frontiers of Urban and Rural Planning

REVIEW ARTICLE

Open Access

City models: past, present and future prospects

Helge Ritter[2], Otthein Herzog[1,4*], Kurt Rothermel[3], Anthony G. Cohn[5,6] and Zhiqiang Wu[1]

Abstract

This paper attempts to take a comprehensive look at the challenges of representing the spatio-temporal structures and dynamic processes that define a city's overall characteristics. For the task of urban planning and urban operation, we take the stance that even if the necessary representations of these structures and processes can be achieved, the most important representation of the relevant mindsets of the citizens are, unfortunately, mostly neglected. After a review of major "traditional" urban models of structures behind urban scale, form, and dynamics, we turn to major recent modeling approaches triggered by recent advances in AI that enable multimodal generative models. Some of these models can create representations of geometries, networks and images, and reason flexibly at a human-compatible semantic level. They provide huge amounts of knowledge extracted from huge collections of text and image documents and cover the required rich representation spectrum including geographic knowledge by different knowledge sources, degrees of granularity and scales. We then discuss what these new opportunities mean for coping with the modeling challenges posed by cities, in particular with regard to the role and impact of citizens and their interactions within the city infrastructure. We propose to integrate these possibilities with existing approaches, such as agent-based models, which opens up new modeling spaces including rich citizen models which are able to also represent social interactions. Finally, we put forward some thoughts about a vision of a "social AI in a city ecosystem" that adds relevant citizen models to state-of-the-art structural and process models. This extended city representation will enable urban planners to establish citizen-oriented planning of city infrastructures, to make them into inviting environments that reconcile and foster human culture, city resilience and sustainability.

1 Introduction: modeling cities – a multifaceted challenge

The complexity of cities makes their modeling into a multifaceted challenge, arising from the numerous perspectives to be covered by models.

Historically, the perhaps earliest city models started with pictures or crude maps to aid finding one's way through the city. With increasing size and complexity of cities, along with improved technical means, city maps improved steadily, with the first printed maps of major cities such as Rome, Antwerp, and Paris appearing in Europe between the 15 th and 16 th centuries, with printed city maps in China dating back at least to the 12 th century, with maps carved in stone even older (Cheng 2022). These may be considered as the earliest prototypes of spatial city models that already also cared about visualization of and semantic information about prominent landmarks, addressing in their essence very similar questions as the much more detailed city models of today (Fig. 1).

Modern cartography had its beginnings around the 18 th century, followed by major advances, such as

*Correspondence:
Otthein Herzog
[1] College of Architecture and Urban Planning, Tongji University, 1239 Siping Road, Shanghai 200092, People's Republic of China
[2] Faculty of Technology, Bielefeld University, Universitätsstr.25, Bielefeld 33615, Germany
[3] Institute of Parallel and Distributed Systems, University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany
[4] FB 03, University of Bremen, Bibliotheksstr. 5, 28539 Bremen, Germany
[5] School of Computer Science, University of Leeds, Leeds LS2 9 JT, UK
[6] The Alan Turing Institute, London NW1 2DB, UK

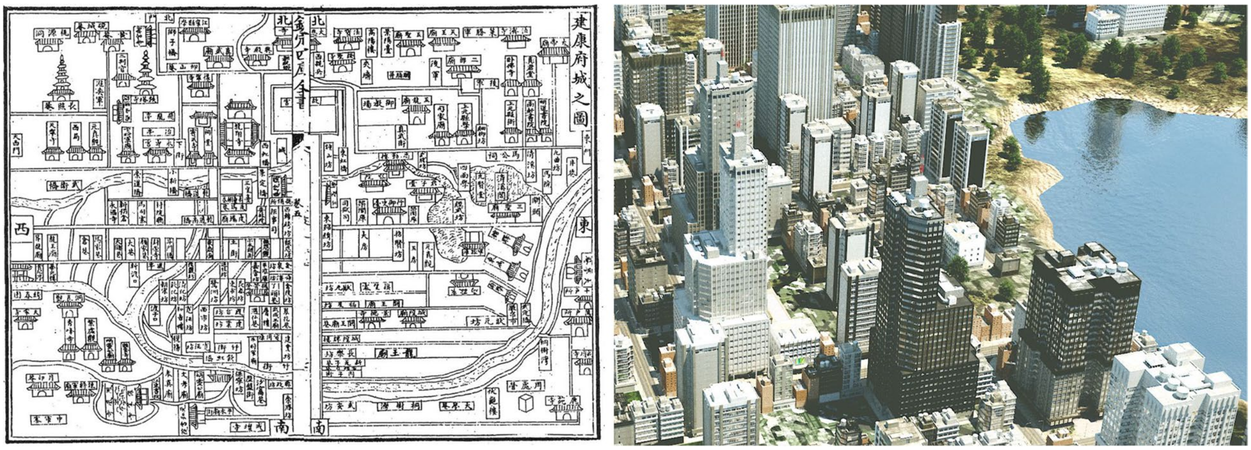

Fig. 1 *Left:* Ancient printed map of the city of Nanjing (around 12 th century (Theobald 2010)). *Right:* 3D model of part of a city model, procedurally generated with Blender (Blender, 2024) and the CitiGen software plugin (Couturier 2024)

triangulation networks and aerial image capture in the 19 th century, and finally the availability of digital computers in the 20 th century, enabling geospatial information systems and the modern digital city maps of our time. The addition of modern Building Information Modeling (BIM) systems towards the turn of the millennium extended these 2D models into fully three-dimensional city models.

The early city models were essentially *static* and focused on *modeling shape* (but maps of other features, such as weather maps, or of the spatial distribution of a wide range of entities, including population, did exist as well). However, these models were purely descriptive and always focused on a single particular instance (such as a map of Rome). However, visualization triggers for more flexibility, e.g., to realize different views with regard to perspective, scale, illumination, weather conditions or seasons in a year. All this requires a model that can render or *simulate a variety of views* under different conditions. Today, such view simulation can happen instantly, including simulated changes to the geometry or even presence of entire buildings or other structures. We owe such models to our reasonably accurate knowledge of the physical laws that are involved when surfaces reflect light to form a picture for an observer, along with powerful computing hardware to simulate the imaging process.

We also may wish models that reflect other properties of cities, such as traffic flows, resilience against flooding or heat, pedestrian flows, and many others. Such models help us *explore consequences changing the city itself*, for instance, by generating alternative plausible city variants "of a certain kind" (e.g., a mid-sized town in a particular region with a particular mix of building styles for its houses). This includes *making predictions*, e.g., generating probable development paths of a given (true or hypothetical) city in the future. Finally, modeling variability can help us synthesize developmental patterns of an entire range of cities that share selected similarities at present.

All these modeling directions differ from our initial visualization example by involving processes for which we do not know good law-like representations to implement a simulation. This problem becomes even more acute when we turn to the impact of *the citizens* since the actions of the citizens are a major, if not the strongest, driver of temporal change. Citizens act in very complex and often highly interdependent ways, giving rise to complex patterns of change at many different time scales, from hours to decades. Although some aspects of their actions may admit aggregated models based on average behavior, there are always macroscopic changes in a city that have originated from decisions of only few or even single individuals: the physical infrastructure is much simpler than the main drivers of its dynamics, the citizens.

This paper will first offer (Section 2) a view on the landscape of existing models, along their path from largely static geometry models to dynamic models, with an emphasis on major ideas, such as fractal geometry, cellular automata, dynamical systems, digital twins and multi-agent models, and how these attempts led to accounting for the important role of citizens. Then it will shift the focus to generative models and recent advances of AI that have enabled their construction in a purely data-driven way (Section 3). Our emphasis will be on the capabilities of models that can generate complex data objects, such as images or other spatial structures, and, particularly, on Large Language Models (LLMs). These models can learn, capture huge amounts of knowledge and bring it to bear on solving tasks that can be specified in natural language, or even in multimodal ways that admit combinations with images and sound. Finally, we will look at the new opportunities and prospects (Section 3) that arise from these innovations and present some thoughts on a vision of a "social AI in city ecosystems" that will enable urban planners to establish citizen-oriented planning of city infrastructures and to make them into inviting environments that reconcile and foster human culture, city resilience and sustainability.

2 The landscape of urban models: scale, form and dynamics

The complexity of cities inevitably calls for many models, each one addressing a subset of features that contribute to characterizing a city. These models will come in many variants, but they all somehow have to say something about how cities vary with scale, in spatial form, and how they vary in time. Moreover, what they have to say can be represented at different levels of abstraction or granularity.

Role of scales. Cities and their structures span across many scales, both in the spatial (meters to kilometers) as well as in the time dimension (minutes to hours, days and months). Spatial and temporal scales are linked in many ways, sometimes following simple patterns, such as changes to large structures usually taking more time to happen.

Complex couplings arise through the spatial coupling of moving entities, such as pedestrians and cars. These couplings and the time scales of their dynamics can vary abruptly when certain "order parameters", such as density, or cross critical thresholds are met (Helbing 2001). Such "emergent" or "critical phenomena" can make it hard or impossible to confine a model to a single time scale and instead one must strive for models that can capture such critical phenomena at the "edge of chaos" (Chen 2009). Moreover, these conditions are not exceptional but a frequent characteristic of complex systems (Siegenfeld and Bar-Yam 2020) and actually prerequisites for important capabilities, such as resilience and innovation. Many elements of a city have memory-like properties, which range from hysteresis over simple adaptation to sophisticated planning and anticipation, which can act in complex, stabilizing or de-stabilizing ways, e.g., when the decision behavior of citizens comes into play, and extend to very long time scales, e.g., due to citizens' ability to create and pursue long-term plans, or slow processes such as demographic change.

Scaling laws express important aspects of scale. With regard to cities, properties such as area, energy consumption, or various connectivity measures (e.g., total road length), necessarily vary as a function of city size (e.g., measured by the number p of inhabitants). It turns out that this variation often takes the form of a *power law* for the quantity of interest: $A = p^d$, where for many cases, the expected values for the exponents d deviate from their "naive" expectations. These deviations often indicate *efficiency gains*, e.g., sublinear ($d < 1$) growth behavior for quantities such as city area or total road length (Molinero and Thurner 2021; Chen 2021). The resulting, usually non-integral exponents expose deviations from familiar, 1D or 2D geometries of road networks or occupied city area, and indicate linkages between scaling laws and models of *urban form* (see below).

Models of urban form. Maps primarily embrace the *geometric aspects* of a city, in the form of 2D, 3D or fractal models. Models can be further distinguished by their focus on *land usage* (2D areas), *connectivity* (graph-based representations), *buildings* (3D structures) or combinations thereof (e.g., planar maps with street networks and embedded areas). These models differ in their extent, e.g., buildings are very localized whereas streets or other transportation facilities can extend across wide ranges.

Statistical models can simplify local detail below a particular scale into averaged local bulk properties, e.g., of city patches. This can make shape properties, such as, e.g., distance or area, scale dependent. Fractal models (see below) offer an elegant solution to represent such nested-ness of detail.

Interaction and proximity. Interactions between citizens are crucial for any city and are usually sensitive to proximity. Land features such as ridges, rivers or lakes, or man-made structures such as fences and building complexes can impact on travel times, effectively "distorting" geometric distance (Barnsley 2014)). Urban road and transportation networks add further reshaping effects on spatio-temporal characteristics of interactions. The resulting complex intertwining of many factors leads to significant deviations of proximity from what would be expected according to Euclidean geometry.

Graph- and network-models. With the advent of wireless networks, which complement physical proximity with a completely new layer of "virtual" proximity, network and graph models, cities have become augmented with an extremely important new layer of transport structures that needs to be included in urban modeling.

Graph and network models offer a unified approach to model both the physical as well as the information transport networks in this layer as general connectivity structures, with geometric analogs to distance, connectedness, and centrality. Because of their discretized nature, graphs are also very suitable to represent *topological properties*, such as connectivity, which are more coarse-grained than geometry. Graphs with weighted edges can also conveniently represent travel distances or times instead of Euclidean distance (Sun et al. 2012). Using such models can avoid overfitting while still representing important properties, such as reachability or clusterings.

Fractal models permit a unified treatment of urban form, proximity structures and scaling properties (Batty and Longley 1986; Batty and Xie 1996; Chen 2021). Moreover, these models have been observed to also capture intriguing aspects of the *growth of cities* (Batty et al. 1989; Shen 2002; Zhang et al. 2022). Fractal geometry can be characterized in several ways, for example by *fractal dimensions* which describe how some size measure S (such as length, area, volume) associated with a geometric object behaves under a linear change of scale. This is described by a power law $S(L) \propto L^D$ where L denotes the scaling factor and D> 0 is the dimension that characterizes the geometry of the object. For fractal objects, this dimension can deviate from the usual integer values.

Besides area, other size measures, such as the total length of road or transportation networks, or the length of coast lines, also exhibit fractal scaling properties indicated by non-integer D dimensions. As a result, the concept of fractal geometry has been found to be useful for characterizing the scaling properties of a variety of urban structures (Chen 2021; Jahanmiri and Parker 2022).

Self-similarity and procedural city models. Fractal scaling behavior is closely linked to the property of *self-similarity* – i.e., when parts of a shape resemble a scaled version of itself or some part of it. This can be turned around, creating a shape from suitably (either deterministically or stochastically) re-arranged copies of a number of initial "generators", and repeating this procedure recursively. Such constructions can be shown to lead to geometries with fractal dimensions that can be directly computed from their recursive construction process. They turn out to offer approximations to many shapes that are encountered in nature (Barnsley 2014; Mandelbrot 1983), as well as a number of man-made artefacts (buildings, cities, road networks). There are different mathematical frameworks to formalize the generative process, such as Lindenmayer- or L-Systems (Prusinkiewicz 1986) or iterated function systems (Barnsley 2014). They triggered the seminal work of Parish and Müller (2001) which led to increasingly refined *procedural city models* (Kelly and McCabe 2006; Vanegas et al. 2009; Kim et al. 2018), and sophisticated city modeling engines, such as CityEngine (Roumpani 2022) which has become widely available in the commercial ESRI software (ESRI 2024). Starting from a given geographical area with land features and additional constraints, such as targeted population density, these models can be parametrized to generate house, land use, and street patterns that cover the given area with an arrangement of houses and streets that follows the prescribed constraints. In a subsequent step, the 2D geometry plan can be augmented with further features, such as building heights and texture choices for walls and roofs (an illustrative result is depicted in Fig. 1 (right)).

Qualitative Spatial Representations (QSR) Cohn and Renz (2008); Chen et al. (2015) is a well developed field which is concerned with the representation of qualitative spatial information and reasoning with it. In natural language, spatial information is usually represented qualitatively (using prepositions such as *on, in, left of, part of, under, touching, east of,...*) and many formalisms have been developed to represent such information. Each of these calculi consists of a set of small finite set of relations – e.g., one well known calculus (RCC) for handling topological information consists of eight *jointly exhaustive and pairwise disjoint* (JEPD) *base* relations allowing one to specify whether two regions are disconnected, touching, partially overlapping, equal, or part of the other one (either touching at the boundary or not); RCC has been implemented in GeoSPARQL. Other calculi focus on representing information about shape, size, directions and may also explicitly handle the case of regions with vague boundaries which is a common occurrence in geographic situations (e.g., exactly where is "downtown"?). Each calculus has associated specialized inference mechanisms, in particular an efficient way to compute the *composition* of two relations – e.g., if one knows that region x is part of region y and region y is part of region z, then one can infer that x must be part of z. Using composition one can not only infer relationships not explicitly in a knowledge base, but also check if a database of facts concerning a set of spatial entities is consistent or not – this might not be the case if the information has been provided by different agents or sensors, or at different times, for example.

Spatial economics. Another class of models employs metric spaces with a continuous distance metric to represent interactions in cities with partial differential equations, similar to field equations in physics. Taking further guidance from the idea of *gravitation* choices for the placement of cities or economic entities are modeled as if seeking minima of potential energy in a fictitious, gravitation-like field of attractive forces between entities for which proximity has an economically favorable effect. This has given rise to spatial models of urban economics that offer an attractive semi-analytic middle ground between highly abstract conceptual and extremely detailed simulation models (Redding and Rossi-Hansberg 2017).

Other spatial models. Spatial cluster structures are the focus of the *space syntax approach* to city modeling (Hillier et al. 1976; Van Nes and Yamu 2021) which is a family of techniques for representing and analysing spatial relationships, especially in an urban context – it views cities as networks of spatial entities formed by the placing, grouping and orientation of buildings, enabling analyses of how streets interact spatially with all the other streets in a particular urban setting. It can be used to understand and predict the effects of spatial configurations on social, economic, and environmental behaviors within urban environments. By examining how streets, buildings, and public spaces are interconnected, space syntax provides insights into the movement patterns, accessibility, and spatial usage within a city. It thus has the potential to help urban planners and architects design more efficient, navigable, and sustainable urban environments by revealing the underlying spatial structure that influences human behaviour. A historically prominent example is *central place theory* (Chen and Zhou 2006) that attempts to capture urban structure in a hierarchical set of *central places* and their neighborhood relationships.

City development and dynamics has been a longstanding concern of urban researchers. *Diffusion-limited aggregation (DLA)* (Witten and Sander 1983; Halsey 2000) is an early example of a model that takes the degree of simplification to the extreme. It models a growth dynamics arising from a large number of particles moving along the random paths of Brownian motion.

This dynamics causes an initial seed to grow into a dendritic pattern that can mimic the development of fractal city shapes over time (Fotheringham et al. 1989; Batty et al. 1989). This has triggered numerous studies of self-organized models for growth processes in nature for their capability to also mimic dynamical processes observed in the development of urban structures (Xu et al. 2007, 2021).

Early dynamical city models. Dynamical city modeling originated with the classical Urban Dynamics model (Forrester 1970; Moody 1970). It uses a parameterized urban space of fixed size with a population which distinguishes managers, skilled and unskilled workers. These seek dwellings in three types of city areas and get their wages from a number of firms categorized into new, mature and old firms, reflecting the different needs of the different worker types. The model is not a spatial model in that it uses aggregated quantities for which it sets up rate equations for their temporal changes. The rate of these changes is specified with equations inspired by classical engineering control systems, emphasizing the important role of closed feedback loops. Although a major advance at its time, many of the limitations of this model were already noted by Forrester himself, amongst the most severe perhaps that the city cannot expand and that the action response patterns of the agents are taken as constant over an entire multi-year simulation time span.

Cellular Automata (CA) models provide an elegant and intuitive approach to connect dynamics with spatial structure. Popularized with Conway's "Game of Life", Tobler (1979) then brought the framework to geographic modeling. Space is modeled as a grid (or a more general type of graph) of cells. Each cell holds a "state variable" describing local properties (such as type of land use, population density etc.) of the corresponding land patch at a particular point in time. A "transition rule" then specifies, for each cell, its (possibly altered) state at the successive time step, using only the current state of the cell, and the states of cells within a certain spatial neighborhood. Applying this rule to all cells iteratively creates the dynamics of the system. The remarkable complexity of the spatio-temporal patterns that can arise even from seemingly simple rules and neighborhoods have led to a wide adoption of CA models for geographic and urban modeling (White and Engelen 2000; Isinkaralar and Varol 2023). Identifying the rule application with decisions of some "agent" connects CA models with the wider class of agent based models.

A famous example is the *Schelling model* and its subsequent variants (Jensen 2022). The model consists of two populations of agents (e.g., "white" and "black") that make repeated decisions on whether to stay in their current location or move to a different vacant place in the grid, with this decision only dependent on the agent distribution in their neighborhood. When the considered alternative place has a more preferable (as defined by a utility function) neighborhood, they move; otherwise they stay. Simulating this model has consistently shown that the model converges to a self-organized demixing of the population into homogeneous domains. Subsequent studies allow for more realism (such as including market mechanisms for house prices, non-symmetric preference functions, bounded rationality of agents, and generalization to more than two-component ethnic mixtures) have reproduced this type of behavior for a wide range of settings (Clark and Fossett 2008).

Dynamical models of transportation systems address another key urban structure. Travel demand models aim to predict the response of transportation demand to modifications of the parameters of the transportation system and the travelers (Bhat and Koppelman 1999). Trip-based travel demand models, which were developed in the late 1950's, used individual trips as the unit of analysis. They assumed that it was possible to predict travel behavior based on household data averaged over a zone. Activity-based travel demand models developed in the late 1970 s understand travel as a derived demand from the need to pursue activity distributed in space (Axhausen and Gärling 1992) and hence focus on household and personal activity scheduling rather than trips isolated from human behavior. In recent years, deep learning approaches have been proposed for both short- and long-term traffic forecasting (Yin et al. 2022; Jiang and Luo 2022). These data-driven methods mine the complex non-linear spatial-temporal patterns of traffic data to make traffic predictions. However, such data is only collected infrequently and expensively from traffic surveys resulting in out-of-date predictions, although call records from cellular communication networks may provide an alternative (Yin et al. 2018; Wang et al. 2024b).

Agent-based models employ mobile agents located in space in order to arrive at finer-grained models that can address spatio-temporal urban and land use phenomena (Huang et al. 2014; Zhang 2021; Crooks et al. 2021). The simulated agents are endowed with some form of behavior that has to be specified, ranging from extremely simple to highly sophisticated and only limited by the computational resources of the simulator and the imagination/creativity of the modeler.

Typically, these agent-based models exploit Cellular Automata, where agents can change between a small number of states and are fixed or limited to simple moves on some discrete grid. Their state changes follow a set of rules that mimic interactions with a (usually) local neighborhood. The choice of neighborhoods and agent rules can vary widely and is critical for the performance of the model. While some models, such as the aforementioned Schelling model, create rules from the intuition of the researcher, models intended for quantitative predictions need rules carefully tuned to data. This, and also to avoid any bias from "handcrafting" rules, requires the construction of agents in a more data driven way (e.g. using regression approaches, "population generation methods" (Sun and Erath 2015) or deep learning).

Compared to human behavior (which includes adaptivity, memory, sensorimotor skills, social interaction, creativity and more), even the classical agent models still appear strongly simplified. However, such simplification can also be designed systematically in order to investigate which behavioral "components" (e.g., automatic obstacle avoidance when moving) contribute to which aspects of city dynamics (e.g., efficient pedestrian stream dynamics).

Digital Twins (DTs) have been developed and successfully used as a comprehensive planning, analysis, and control concept for complex cyber-physical systems, e.g., in manufacturing (Mihai et al. 2022). Subsequently, the same DT approach was also taken up for city planning (Ferré-Bigorra et al. 2022), another complex area where it is of utmost importance to avoid costly errors and model the (future) reality as closely as possible, despite the numerous interrelationships between the many different aspects of a city (Herzog et al. 2023). In addition, these model DTs must be capable to model societal necessities such as an emphasis on sustainability and reaching relevant UN Sustainability Development Goals (SDGs) and to be updatable with not only new data but also new functionalities (Herzog et al. 2023).

While Digital Twins offer a big step towards agents with rich behavior based on a sophisticated model, there are still major gaps when it comes to a comparison with real humans populating an urban environment, and LLMs offer a potential solution to help bridge such gaps. Wang et al. (2024b) surveys the literature on integrating LLMs and agent technology; several of these approaches (e.g., BabyAGI, CAMEL, Multi-Agent Debate, MetaGPT) are based on multi-agent systems that support collaboration between agents to accomplish a common task. AutoGen (Wu et al. 2023) is an open source framework for building multi-agent systems, where agents can integrate humans, tools, and LLMs, or combinations thereof. Agents can use many of the capabilities of advanced LLMs, such as role-playing, state inference, making progress based on conversation history, providing feedback, adapting from feedback, and coding.

The combination of LLMs and agent technology opens up a wide range of promising applications. For example, LLMs can be used to decompose potentially complex tasks into simpler subtasks organized as a workflow. To execute the subtasks, the LLM generates code that is passed to agents, which have the ability to execute tools through code execution (Li and Ning 2023; Mansourian and Oucheikh 2024). In addition, agents and LLMs can work together to solve potentially complex optimization or simulation problems, where the LLM generates solver code based on a natural language query and translates the solver result into a natural language response (Li et al. 2023a). Finally, an agent can assist urban planners in performing complex tasks such as text review, auditing, or evaluation, and assist them in tasks such as reverse geocoding, knowledge graph construction, and image captioning (Zhu et al. 2024a).

3 What AI generative models can do for city models

Advances in AI and machine learning have greatly impacted on our modeling tools in general. Before these developments, our only way to create models relied on careful manual design. This usually required significant insights into the to-be-modeled processes, ideally distilled into mathematical equations to provide the basis for a computational model. The role of data was limited to identify suitable settings of the (usually few) parameters of the model to make it fit given data and to predict new data about the phenomenon of interest. The flexibility of these parameterized models was severely limited by the relatively low number of their parameters. Increasing this flexibility by introducing more parameters usually led to overfitting and poor performance on new data. On the other hand, for processes whose inner structure or basic laws were basically unknown, models with such limited flexibility were very difficult or even impossible to adapt to fit the empirical data. As it is well known, this changed around a decade ago when it was discovered (in the context of models for object recognition from images (Krizhevsky et al. 2012)[1]) that deep neural networks did not pose this overfitting problem of traditional parameterized models, opening the avenue to what is now known as *data-driven modeling*. In this approach, the model emerges almost completely[2] as the result of training a deep neural network on a training data set, where usually both the deep neural network and the training data set must be extremely large to make the method work.

This revolution has greatly pushed forward the frontiers of modeling in many fields, and notably in urban science among them. Here (as well as in many other fields), a particular impact comes from the important subclass of *generative AI models* that can generate complex data objects, such as images, texts, or also almost any other kind of complex data item, provided that training data sets can provide enough examples.

3.1 Generative models for images, 3D structures and maps

Many aspects of urban modeling are related to visualization, a process governed by well-known physics. However, details of surface properties that impact appearance remain cumbersome to model. Here, deep learning models have been able to substitute physics-based modeling with purely data-driven visualization models that learn from large image databases. Most of these models originate from Generative Adversarial Networks (GANs, Goodfellow et al. (2014)).

GANs provide a very versatile class of data-driven generative models for images (and other high-dimensional objects) that are generated from having two networks that learn in an opposing ("adversarial") way from given data (such as an image data base): a first network (the final generator) transforms its input (in the simplest case some random vector) into a data sample (e.g., an image). Its learning task is to respond to each input with a sample that resembles the samples (e.g., images) in the data base. A second network, the discriminator, receives samples randomly chosen from responses of the generator or from the database and has to learn the binary decision whether they stem from the former or from the latter. Failing to make the right decision is used as a reward for the generator, thereby driving its learning towards the production of samples that become more and more indistinguishable from true database elements.

This technique has now become an almost standard way to enable the generation of complex data items (of which images are just one major case) from many domains, without requiring any explicit specification of the desired result except for providing a large database of examples (Hong et al. 2019; Gui et al. 2020). By conditioning the learning of the two subnets on a shared steering input, the output can be controlled in a flexible manner. This allows transformations between vastly differing representations of data items to be implemented. For example, the control input can be image sketches or stylized facade schemas (Isola et al. 2017; Bachl and Ferreira 2019), which the generator then transforms into photorealistic images. Other recent works demonstrate the generation of 2D designs of house floor plans ("House-GAN", Nauata et al. (2020)) or entire 3D structures for buildings ("Building GAN", Chang et al. (2021)) from simple text-based specifications. These techniques have been elaborated in Quan (2022), allowing even non-experts to generate entire city street layouts that mimic the urban form of a provided example.

Recently, another generative approach, *diffusion models*, has emerged that can produce similar results. It is

[1] A milestone paper with more than 130000 citations

[2] The only handcrafted specifications being the choices of the neural network architecture and the training data set plus a handful of hyperparameters.

based on a different principle, exploiting the computational inversion of a diffusion process. The idea is to generate training sequences, each of which is a sequence of progressively noisy versions of some different starting image, with each sequence ending when the final image can no longer be distinguished from pure noise. The generator is obtained by training a predictor network on the reversed sequences. This makes the trained predictor able to iteratively transform even pure noise into an image from the image space used to construct the training sequences (Song et al. 2020). Like GANs, this approach has become extended with conditioning inputs and further refinements and has begun to replace GAN generators due to its potential for even higher image quality (Dhariwal and Nichol 2021) – for a survey see Yang et al. (2023). Again, it has become combined with front-ends for flexible text-based control of image generation, thereby often drawing on large language models (LLMs) as pioneered by GPT and its successors (see below). These approaches are already available in a multitude of commercial (e.g., Midjourney (2024)) or free (e.g., Stable Diffusion) platforms with convenient text-based interfaces to make the image generation process also feasible for non-experts (see Fig. 2 for an example). It is important to note that these results come without any guarantees about physical feasibility since the models were only trained with visual images, not with any deeper representations to capture static or other physical laws.

Fig. 2 Futuristic city rendered with Midjourney using a specific text prompt (Imagine you are an urban designer. Create a view of a city with a harmonic balance between architecture and green zones. The buildings should use curved surfaces and reflect a sharing of style without being monotonous. Use building smooth, curved forms that blend well with forms of nature. Make the buildings stand out aesthetically by giving them bright colors. Render the cityscape in mild late afternoon sunlight. Avoid skyscrapers and plain rectangularity. Pay careful attention to each of the requirements)

Besides visualization, data-driven approaches do advance into numerous other tasks that are important for urban modeling, including generation of land use patterns from surrounding spatial context to aid urban planning (Wang et al. 2023a), building layout generation (Jiang et al. 2023), wind flow modeling (Kastner and Dogan 2023), floor plan design from human activity patterns in buildings (Wang et al. 2021), lane extraction from aerial images (Ruiz et al. 2024), or generative design of walkable cool areas based on GAN-generated heat patterns (Li et al. 2024b). Other works demonstrate the design of complex "citiscapes" populated with 3D building models, streets and lakes from aerial images (Zhang et al. 2020), or the "CityDreamer" (Xie et al. 2024) that can generate spatially unbound 3D cities including the rendering of visual appearances for their 3D-buildings.

3.2 Large language models

Large Language Models (LLMs) (Devlin et al. 2019; Brown et al. 2020) are a form of so called *generative AI*. They are neural networks with billions of parameters (the weights of the interconnections between nodes in the network). These parameters are set by training the model of very large amounts of text to predict what might come next given any prompt. Although simple next word prediction seems a very primitive facility, in fact LLMs have displayed remarkable fluency and ability to answer seemly any question on any topic. They are sometimes termed "foundation models" since they can potentially be used as the foundation for an AI system in virtually any domain or application area, perhaps just requiring further "fine tuning" or careful prompting strategy (i.e., the precise wording and form of question given to the LLM). The performance of LLMs has improved dramatically in recent years with many claims made about their abilities including their ability to reason (Creswell and Shanahan 2022; Huang and Chang 2023; Kojima et al. 2022), for commonsense (Belussi et al. 2022), and their ability to pass tests designed to test human ability, many point to, possibly fundamental, limitations in their abilities, and suggest that they alone will not lead to Artificial General Intelligence (AGI) (Baum 2017), but that some kind of neuro-symbolic AI (d'Avila Garcez and Lamb 2023) might be a more likely route to AGI; it has already been shown that, for example using LLMs to translate a problem into a logical formulation and then using an external reasoner can improve performance (e.g. Li et al. (2024a)). The problem addressed in Li et al. (2024a) was a spatial one based on the StepGame benchmark dataset. The native LLM performance was less good than when the external reasoner was employed in conjunction with the LLM. There have been other demonstrations of the difficulties that LLMs have with spatial reasoning, including the difficulty of reasoning about cardinal directions, e.g., Cohn and Blackwell (2024), particularly intercardinal directions (NE, SW, NW, SE).

A more recent development is the emergence of multimodal LLMs (MLLMs) – which can for example, process and generate images as well as just text (Zhang et al. 2024). A recent survey is Zhang et al. (2024). In principle, multi-modal LLMs include other kinds of modalities, for example speech input and generation, but we will just consider vision-language MLLMs here. There are two aspects to consider: (1) MLLMs which can generate images given text prompt and (2) MLLMs which can accept images or other kinds of visual inputs and analyse them, and then respond to textual prompts. We have already seen the impressive abilities of systems such as Midjourney and DALL-E to generate images to order in Fig. 2. However, their map drawing abilities are very poor – see, for example, Fig. 3 – there are so many things wrong with the map including the fact that the requested route is not shown that one does not know where to start.

MLLM's capabilities regarding the analysis of images and maps in particular have been improving. For example, consider the map displayed in Fig. 4 and the descriptions generated by ChatGPT- 4o which can be found in the Appendix as Fig. A1. The description is generally very good and accurate, though it contains a few subjective opinions or surmises about the region not backed up by the image (e.g., the School being an important community institution, and that the bridge is a crucial piece of infrastructure) – both are in fact true but cannot be inferred directly from the map – though it is possible the LLM used information gathered during training to make these statements.

However, when it comes to interpreting sketch maps LLMs seem to struggle more – see Fig. 5 and the description given of it in the Appendix, Fig. A3. Some parts of the description are good, in particular all the text relating to the roads – except for the purple road which is ignored in much of the output. However, the locations of the buildings are mostly wrong. Moreover, if asked to plot a route from the Fishmonger to the School, many mistakes are made (e.g., one should not go past the newsagent, one should turn right not left on to the red road – see Fig. A2 in the Appendix).

3.3 Geo-LLMs

In the previous section we looked at LLMs generally and their abilities to reason with or generate spatial information and maps. In this subsection we report on some of the growing body of research into geo-LLMs. For example, GEOLM (Li et al. 2023b) aims to connect geo-entity mentions in text to geospatial information extracted from geographical databases; the authors demonstrate

Fig. 3 The image drawn on 3/9/24 by ChatGPT- 4o to the prompt "give me an image of a map of the sea route from Southampton to Barcelona."

that GEOLM has some capability to support toponym recognition, toponym linking, relation extraction, and geo-entity typing. Qi et al. (2023) propose that rather than using LLMs *for* spatial databases, they should be used *as* spatial databases – "Machine Learning Models as a Spatial Database" (MaaSDB). Whilst they point out the advantages this might bring (learning from and integrating information on both structured and unstructured data; full integration between the natural language interface and the data itself), this is not without risks – LLMs are well known for their propensity to "hallucinate" (Huang et al. 2023) and thus relying on an LLM for accurate spatial fact retrieval may yield incorrect results (and indeed the paper itself gives examples of incorrect information being returned in a small number of cases in their preliminary study using ChatGPT). Other work investigating geo-spatial capabilities of LLMs includes (Bhandari et al. 2023) who conclude that at least the largest LLMs, and those that have been instruction-tuned do have fair abilities.

Fig. 4 A map of Pool-in-Wharfedale in West Yorkshire, UK. Presented to ChatGPT- 4o as a raster image taken from Open Street Map. The given prompt on 4 Sept 2024 was: "Consider the attached map. Please describe the main geographical and built environment features"

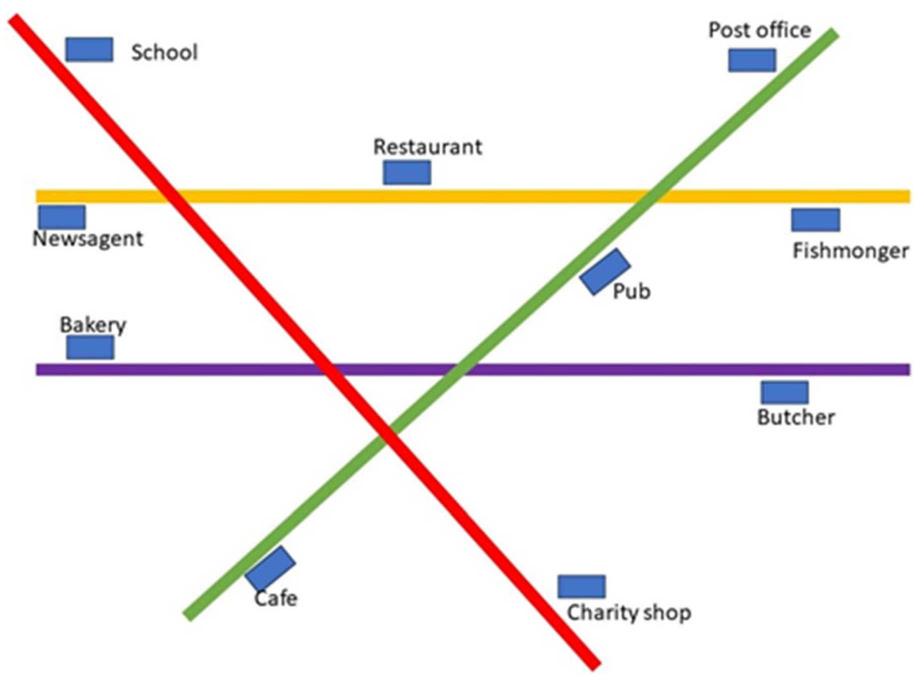

Fig. 5 A sketch map given to ChatGPT4-o on 3rd September 2024 with this instruction: "Please consider the attached sketch map and describe the various features on it in sufficient detail that someone else could create a reasonable copy of it." See Fig. A3 in the Appendix for the LLM's response

An interesting approach to interacting with geographic data extracted from a map is presented by Unlu (2023). In this work, a number of points were chosen such that 300 metre radius circles centred on these points were densely tagged with features (e.g. schools, cafes, ATMs...). The amenities of each area are then represented as a *preprompt*. E.g.:

> *This is a circular area of radius of 300 meters that intersects province(s) of İstanbul and district(s) of Fatih. There are 3 atm(s), 2 bank(s), 1 bureau_de_change(s), 18 cafe(s), 2 clinic(s), 1 court_house(s), 2 dentist(s), 1 driving_school(s), 2 events_venue(s), 11 fast_food(s), 1 guest_house(s), 3 hospital(s), 11 parking(s), 33 pharmacy(s), 9 place_of_worship(s), 1 post_office(s), 43 restaurant(s), 5 school(s), 1 shower(s). There are 525 buildings which cover 31% of the total area. It contains 289 meters of platform rail, 100 meters of footway road, 80 meters of pedestrian road, 44 meters of primary_link road, 2786 meters of residential road, 283 meters of service road, 20 meters of steps road, 1005 meters of tertiary road, 62 meters of tertiary_link road, 249 meters of unclassified road.*

ChatGPT 3.5 Turbo was then used to generate prompt answer pairs from these preprompts, using prompts such as:

> *I will give these types of preprompts and you will generate prompt-answer pairs in python list of dictionaries format. These prompts should be questions that businessmen, citizens, tourists would demand based on the data in the preprompt. Generate 50 prompt-answer pairs with very diverse topics. Important : Do not generate prompts that data in preprompt is not sufficient to answer!*

ChatGPT 3.5 Turbo then produced prompt-answer pairs such as:

> *Question : Tell me about the options for cultural enthusiasts with a gallery nearby.*
> *Answer : Cultural enthusiasts can explore the gallery in this area.*

4111 such prompt-answer pairs (generated from 81 preprompts) were then used to fine-tune an LLM (Falcon 1B RW) using Low Rank Adaption (LORA) (Hu et al. 2021). The resulting fine-tuned model is able to answer questions similar to those in the fine-tuning set about novel regions.

A further experiment reported in the paper is to map the (GLOVE) embeddings of the preprompts onto a "D space using universal manifold approximation and projection (UMAP) – and then assign colours to each embedding based on its UMAP score – these can then be plotted on a map as shown in Fig. 6.

Mai et al. (2023) discuss some possibilities for exploiting Foundation Models in geospatial situations, investigating the performance of a number of different models across seven tasks. They conclude that on toponym recognition and location description recognition (which are purely textual tasks), LLMs can outperform task specific fully supervised models. On the other hand for multi-modal problems, such as land classification from satellite imagery, or street view image-based urban noise intensity classification, specialist models outperformed multi-modal foundation models. They conclude that multi-modal models specifically built or tuned for geospatial data should be an aim for future work.

SpaBERT (Li et al. 2022b) is a spatial language model based on the BERT LLM (Devlin et al. 2019). Just as standard LLMs use embeddings based on neighbouring words in sentences (which SpaBERT continues to use), SpaBERT also builds embeddings based on neighbouring entities in geographic space; these are linearised based on the distance to the *pivot* entity; SpaBERT also builds a third embedding based on the distances to the nearest neighbours in the x and y dimensions separately. SpaBERT is then pretrained in a standard way with masked language modeling and masked entity prediction tasks to learn spatial dependencies. When applied to two downstream tasks: geo-entity typing and geo-entity linking, SpaBERT shows significant performance improvement compared to standard LLMs without any specific spatial competency.

3.4 LLMs and Retrieval Augmented Generation (RAG)

We have already mentioned above the problem of "hallucination" – that LLMs tend to produce text which contains statements which are not true – see Hong et al. (2024) for an open effort to measure the extent to which this happens via a series of benchmarks. There are a number of ways in which this issue might be mitigated, although it is unlikely that hallucinations can be eliminated - indeed (Banerjee et al. 2024) argue that they are something we need to live with, arguing that "every stage of the LLM process – from training data compilation to fact retrieval, intent classification, and text generation – will have a non-zero probability of producing hallucinations". Similarly (Xu et al. 2024b) use results from learning theory to demonstrate that LLMs will always hallucinate. However, there is a community focus on finding ways to mitigate the problem, recently surveyed by Tonmoy et al. (2024). They outline a taxonomy of possible methods, with the two top-level categories being prompt engineering and model

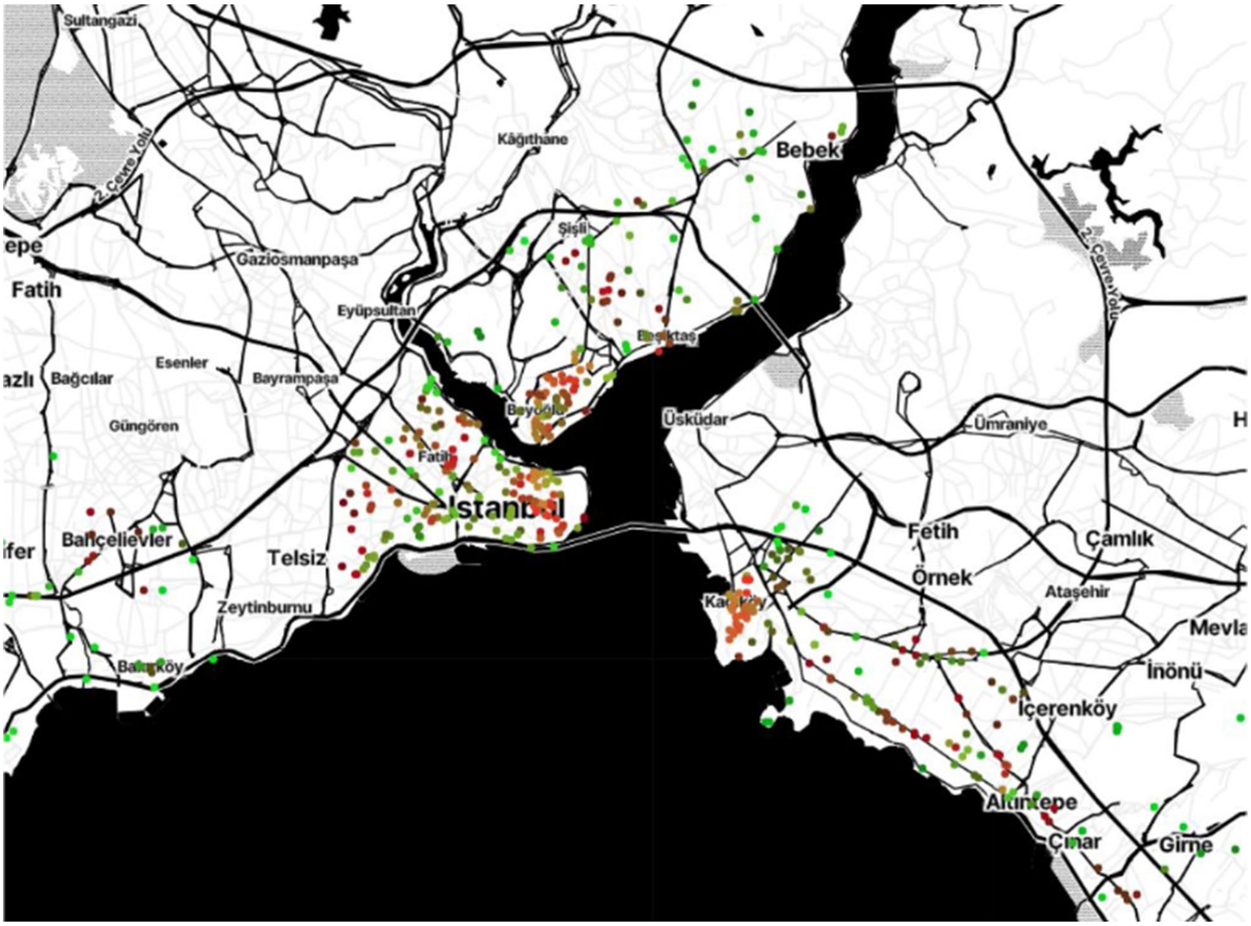

Fig. 6 Embeddings of preprompts visualised on a map where colours are assigned proportionally to UMAP values. "Bright red colors indicate more touristic locations, dark red colors more indicate business/commercial districts and bright greenish colors indicate relatively empty spaces, residential areas" (Unlu 2023, Fig. 4)

development. Within the former category, Prompt Tuning, Self-Refinement and RAG form the next level of the classification. RAG is of particular interest in our domain of interest. RAG addresses one of LLM's biggest weak spots – since LLMs are only (re)-trained rather infrequently – because of the huge cost – so their "knowledge" may date quickly, particularly about propositions whose truth may change frequently. RAG provides a way for LLMs to tap into authoritative symbolic knowledge bases as a secondary explicit knowledge source rather than relying on information held implicitly in its model from possibly outdated training data. This symbolic knowledge can be exploited in three different ways by a RAG enabled system: (1) the external knowledge can be added to the user-supplied prompt before generation starts; (2) during generation, for example by validating generated facts and rectifying them when necessary; (3) waiting until an entire response has been generated before checking it against external knowledge to the LLM and adapting the responses as necessary.

4 The future of urban models: prospects and opportunities

Cities are for citizens. The introduction of agent-based modeling techniques has allowed for significant strides towards more realistic models, but the created agent models were still severely limited by the comparably simple behaviors that could be realized in those agents.

4.1 What is missing?

So far, urban planning has been mainly concerned with the static infrastructure of cities. Adding virtual citizens and allowing them to move and behave in their environment according to their needs is a very important extension. This can be built using the agent-based modeling techniques described above, but it is an open question whether these modeling methods can lead to agents that have sufficient complexity to yield systems with an adequate level of realism. The simple parameterised agent models that aim to be "statistically correct" with regard to their aggregated properties (such as means and correlation functions, see, e.g., Sun and Erath (2015)) can indeed capture some useful aspects of behavior to drive models such as transport systems under loads of simulated passengers. However, these models usually attempt an approximation at a certain time scale, with the agents deriving their behavior from relatively simple sensor signals and following mostly fixed parameterised rules to compute their movement decisions or trajectories. While well calibrated models of that type have shown useful predictive capabilities for real systems, they are limited in many ways, e.g., they cannot deal well with changing conditions, such as unusual weather conditions, emergency situations or simply drifts due to technology changes (e.g., increasing numbers of cars, with changing driving characteristics due to partly or fully autonomous driving systems, changes in car utilization due to sharing or home office). If at all, they have only very simple goals, intentions, emotions, adaptivity or creative capabilities. This exposes a stark difference between current agent models and real citizens.

Citizens have mind states. The behavior of citizens can be extremely rich, reflecting their largely unobservable states of mind that make them extremely adaptive and able to correlate their behavior with memories from decades ago. They enable citizens to shape and take highly informed initiatives, to be emotional, have expectations and values and to be creative. As a result, they can engage in rich interactions among themselves, making reciprocal representations of each other, and interact similarly with the city in which they live. From the observable, outer side, this fills the city with complex interaction patterns that actually underlie most of its development. In addition, the city becomes connected and embedded in the mental spaces of its citizens. In turn, these mental spaces become filled with experiences that arise from the interactions with others and with what the city has to offer to them. This shapes needs, demands, and resources that reshape mind states, changing the complex relational fabric that the citizens weave into the city. It extends the city's physical space with a rich social space in which the citizens' mental spaces are connected and interrelated. While this social space (along with the individual citizens's mental spaces) can only be observed indirectly, by inferring some of their content from their shaping of what the citizens do (e.g., usage of media), it is an entity that is at the root of almost all impacts that the citizens bring to bear on the city.

In the following, we take a brief tour through several of these aspects.

Capturing rich knowledge. Human behaviors obtain their richness from the huge amount of knowledge that affects human choices. This knowledge has a complex and hierarchical structure: the largest part – our semantic memory – has been distilled from lifelong learning from huge numbers of contexts. Another part is from our episodic memory linked to specific episodes in our past. And a third part of our short term and our working memory captures very recent contexts and events just minutes, hours and days ago. Capturing such rich and hierarchical knowledge requires new methods that have opened up powerful ways of creating models completely from data, leading to new types of generative models culminating in the LLMs epitomized by GPT and its successors. The next section is dedicated to providing a compact review of these new possibilities. Finally, we will attempt to look at how this opens up exciting perspectives to complement modern agent models such as Digital Twins with powerful capabilities to bring city models to a new level.

Integration of demands and resources as key drivers. Citizens are both the carriers of resources as well as of demands. A city is attractive when it supports a good matching between resources and citizens' demands. Therefore, we need models that allow for the introduction of resource-demand matchings into the planning process. Such matchings have already been considered as important elements of transportation models (Zhang 2021). Matching processes that are focused on local neighborhoods are – in highly idealized form – also at the core of the cellular Schelling model (Jensen 2022), in which residents match their expectations about the ethnic composition of their neighborhood with possible choices for a picking their residence. This matching model has been shown to explain the strong tendency of ethnic group mixtures toward spatial segregation.

The importance of linking the needs of the citizens with the urban planning processes has been hardly discussed in the literature (Alsayed 2024). González-Méndez et al. (2021) provides one recent example, aiming towards an agent-driven platform that can mediate between different stakeholders in the process.

Smart support environments. The now wide-spread availability of human-compatible perception and natural language capabilities in technical artefacts and the ability to compile huge amounts of human expertise into digital assistants leads to the question how to combine such skills with human expert guidance. This requires to modify standard deep-learning architectures to provide for meaningful inclusion of human expert guidance, e.g., when realising intelligent urban planning assistance functions (Fang et al. 2022). It is also a question to the field of human-computer interaction, extending its scope from artefacts surrounded by humans to entire environments inhabited by humans. For the case of buildings this has spurred human-building-interaction (HBI) as a new branch of HCI (Alavi et al. 2019). As the next logical step we foresee its extension to the even more comprehensive field of human-city-interaction, with the task to elucidate the integration of digital and AI-driven technology with the planning, maintenance and operation of urban structures and services in ways that enhance urban quality and adress human needs.

Citizens' mental spaces. Human interactions are not only highly variable, they also add a level that is not present in physical interactions that are familiar from the physical city infrastructure. This level is constituted of *mental spaces* that are inseparable from human agents. Humans not only feel anchored in the physical space of the city. They also augment it with their mental space which gives them imagination and the power of abstraction. They can populate their mental space with their individual versions of the city or parts thereof, e.g., creating their individual version of the subset of city places which they know and their preferred commuting trails between them (Gibała-Kapecka and Kapecki 2016). They can create mental pictures of "their ideal city" and their life therein, or mental maps of safety zones, recreation areas, landmarks or networks with other citizens. And they can open further spaces accommodating their needs, hobbies, dreams and desires. All these possibilities are ultimately rooted in their experiences in the physical city, but extend its physical space and feed back on it in rich and varied ways (Mehaffy and Elmlund 2023) that reflect their rich relationships with the city they live in. Can we build models that represent aspects of these mental spaces that are shared among large groups of citizens? How would these models interact with the physical city? How do the physical and the shared mental city spaces jointly affect decisions, identity or creativity of the citizens? All these are modeling challenges for the future almost untouched so far.

Social space. Of these mental spaces there is a particularly important one that is specialized on the social dimension of human interactions and, therefore, deserves special attention. The social nature of human interactions becomes already relevant when a human interacts with an artificial agent, such as an avatar in a mobile phone or an embodied robot. The human will locate the agent not only in physical space, but also in social space, perceiving the agent, e.g., as representing an external agency or alternatively as a member of the human's household. This leads to different expectations about agent role, trustability, party taking in decisions, or compatibility of interests and intentions. This requires a "**Social AI**" that extends beyond the merely functional level and that can take social aspects into account in its interactions and services. Conceptually, this again can be framed as the augmentation of the city with another space that represents the social interrelations among the citizens. Here too, we may distinguish between a "shared approximation" of what might be termed "objective social space", and individual spaces (as specialized versions of the individual social spaces of citizens) that represent the individual perspectives of citizens' own social embedding in the city.

Social AI*.* While traditional city models were focused primarily on the functionality of the infrastructure of a city, with subsequently developed agent models adding to that level a layer of behavior in the form of (usually highly simplified) decision-making (including decisions about agent movements, e.g., when the agents model pedestrians or drivers), we are now at the dawn of models that can include rich social behavior, identity and creativity enabled through the presence of these social and further mental spaces. For instance, geotagged twitter data have been used to create models about demographic variations of distancing practices during the COVID- 19 pandemic (Xu et al. 2024a). Commuting efficiency (Ling et al. 2024) quantifies how well urban policies optimize the connection between job and residential locations, and how the degree of optimization varies depending on socio-demographic factors. The path towards smart cities leads to an increasing multitude of platforms to capture data about such and numerous further aspects of citizen life (Pal et al. 2018), providing a rapidly evolving basis for more detailed models of citizen behavior.

This will connect city models with the concept of a "Social AI" as a key element to adequately frame the rich interactions of citizens as social entities in interaction with each other and their mental spaces, and also with AI agents that will become an increasingly pervasive part of the future infrastructure of cities, making cities accommodate a hybrid society of humans and artificial agents which can interact in new synergies to drive the emergence of our cities of the future.

Interdisciplinarity. The above makes clear that a proper addressing of the complexity of urban models necessarily has to cut across many disciplines:

Since cities connect the lives of their citizens, appropriate models have to cope with tasks that involve numerous areas in addition to urban and computer science. To model the forces and dynamics that unfold when citizens meet and move in cities needs to reach out into fields that deal with us humans: psychology, sociology, health sciences, arts and communication, and what it takes to organise us: education, economy, law, and certainly many more. Neither last nor least in this enumeration is ethics: modern cities and their data- and AI-driven models have many of their benefits connected with significant ethical questions concerning privacy, freedom, forms of governance, ownership, values, or norms.

Case studies. Resources and methods for data-driven models and modern AI methods working at the scale of cities are rather recent developments. Thus case studies typically were initially restricted to major metropoles and will certainly become more numerous within the next years. The case study (Sanchez-Sepulveda et al. 2024) for the city of Barcelona provides an early example of the use of data-driven models to inform city planning with a focus on improving walkability, transportation and safety. Another example (Floridi 2020) presents insights and benefits for the cities of Amsterdam and Helsinki when making AI services accessible through centralized AI registers. The case of the city of Singapore (Das and Kwek 2024) illustrates how addressing the needs of the COVID pandemic benefited from AI-driven approaches to realize different subsystems that became gradually more and more integrated and were adapted to serve new tasks after the pandemic, leading to a seamless merging of IoT and AI to explore different ways of ultimately making a city adaptive in real-time, e.g., through smart gantries or police robots perceiving and reacting to human behavior to enforce policy rules. Typically, case studies can only address a subset of the wide range of AI use cases that are relevant to a city. Focusing on the transportation sector, Patil et al. (2024) provides an overview with a taxonomy of major case studies obtained from an extensive literature search on AI applications in urban contexts. A very large use case has been launched around 2017 in the Haidian district of Beijing (Cugurullo and Xu 2024). After a concept phase and the formation of a consortium of firms and governmental institutions the implementation was started in 2020. So far, the project has connected more than 14,000 cameras with a large network of server nodes and two centralized data centers that are running algorithms and AI methods to process the data to provide a layer of automated city management based on combining AI methods for visual recognition and event prediction akin to a "brain" that must run a body that is in this case a city. This approach has inspired an increasing number of similar "city brain" projects for other Chinese cities. Besides a general feasibility of the approach, it has led to the insight that different cities may require different tailored "brain structures" and that adequate participation by the residents remains a major challenge for the current organisational format.

4.2 Realizing rich virtual citizen models

We contend that LLMs put us in a position to create agents that can draw on extensive knowledge for choosing their actions, and in a way that can take complex context into account. Moreover, the nature of LLMs provides these agents with an intrinsic form of short term memory that makes their responses contingent on their interactions within a recent time horizon that be made to include dialog spans of many thousand items if desired. This makes these agents quick to adapt to their recent past, a property which has at best been extremely simple and limited in agents used by previous techniques.

These agents allow city models to be brought to a new level where the city is modeled *together with the interactions with and among its virtual citizens*, which can access the facilities of virtual city models in a knowledgeable and adaptive way.

This will open up entirely new possibilities for urban planners: they can devise different city layouts, populate them with different mixtures of artificial "agent personas" (which can be quite easily obtained with combinations of prompting and fine-tuning techniques), and simulate the resulting dynamics to predict the impact on citizens and explore and evaluate different designs.

First steps towards such models already exist. Wang et al. (2024a) show the realization of an LLM-agent based on GPT- 3.5 that can create personal mobility data that then can be tuned with available human mobility data. They demonstrate that their citizen emulation leads to appropriate activity changes when the agent receives a prompt that informs about the presence of a pandemic and pertinent government advice. Park et al. (2023) show that a group of simulated citizens placed in a small neighborhood of a simulated city area unfold a variety of interesting behaviors in their environment, such as going for a walk, making appointments, meeting at a café, or sharing news with a colleague, demonstrating that LLM agents can model complex social interactions and decision-making processes. Besides enriching models with realistic citizen behaviors LLM agents can also be employed to contribute to a variety of city planning tasks directly, for instance, by using fine-tuning of base models with scientific literature to create experts for specialized areas, such as urban renewal (Wang et al. 2023b).

These possibilities have triggered a significant number of proposals for integrating LLMs with agent technology (Wang et al. 2024b). Some of these proposals are based on single-agent systems, such as AutoGPT or LangChain Agents, while others employ a collection of collaborative agents, such as aBabyAGI, CAMEL, or AutoGen. In the latter case, the agents may incorporate individual LLMs to support collaborative task solving (Wu et al. 2023).

An attractive feature of these LLM integrations in architectures that allow the flexible interfacing of LLMs with additional functionalities (e.g., sensors for perception, databases for specialized information queries, or actuators for allowing physical or virtual actions) is to give the LLM the possibility to *access existing tools* to support operations for which a variety of established solutions of proven quality exist. This permits the LLM to focus on solving problems at a higher level of abstraction, away from the details of the more specialized operations covered by the tool. In essence, LLM support can be limited to combining existing functional building blocks into a workflow to solve the given task.

For instance, as we have seen, current LLMs still have difficulties with many spatial tasks. Thus, basic spatial functions, such as querying a GIS database or visualizing spatial data on a map, can be safely delegated to a tool, such as ArcGIS (ESRI 2024), while the task chain for solving a complex task using these basic functions might be generated by an LLM.

For example, Li et al. (2022a) proposed a single-agent approach for performing GIS tasks. An LLM embedded in an autonomous agent performs the planning task. The agent uses basic GIS functionality such as spatial databases, spatial analysis functions or visualization tools together with existing GIS data to perform its task. The LLM establishes a work flow, generates code to call basic GIS functions and to assemble the final result.

4.3 Towards LLM ecosystems

However, the very high computational and memory demands of the currently best LLMs makes their use within such LLM agent frameworks extremely compute intensive and puts severe limits on the number of agents that can be simulated within reasonable resources. Using these "huge" LLMs to train smaller LLMs has enabled the "distillation" of their knowledge into much smaller footprints, giving rise thus to large number of "small LLMs" (SLMs) (Zhou et al. 2023). Despite their much smaller size the most recent of these SLMs often sacrifice only little of the performance of their much larger brethren. At the same time, they can be tuned (or sometimes even trained from scratch) with much more modest resources. Their execution is usually possible, even with several instances simultaneously, on desktop computers down to laptops or even mobile phones (Chen and Li 2024). Besides clever training schemes and the use of carefully curated data, their compact size has also been enabled through more compute efficient architectures and algorithmic improvements ("weight quantization") that allow them to work with extremely low-resolution parameters replacing standard 32-bit float values for network parameters by coarser number representations that occupy only 3–4 bit, bringing a ten-fold reduction in required memory size. Recently, this has been pushed to the extreme of ternary (−1/0/1) weight and binary (0/1) activity values. With suitably adapted algorithms for training and execution these models only require additions and allow for a very energy parsimonious execution on dedicated hardware (such as standard FPGAs) (Chen et al. 2024).

The availability of such very parsimonious LLMs opens up *the vision of creating systems populated by large numbers of SLMs* to simulate correspondingly large numbers of citizens for a new generation of *bottom-up driven urban models*, that at the same time *are connected to our high-level semantic world*. Such a system could then bring the "mental citizen spaces" into the urban model and simulate the interaction between the physics of the buildings and the mental spaces of the citizens. While such modeling will always be too coarse to obtain predictions about mental spaces of individuals, it may be useful to address phenomena such as the clustering or spreading of opinions and sentiments, their relation to spatial structure (e.g., well-connected vs. scantly connected areas, or local ethnic variations in the city) or to changes in the urban infrastructure (e.g., new apartment and shopping complexes, roads, or park areas).

Alternatively, we can realise intelligent distributed ecosystems that can be *embedded in a city*, with each node providing expertise for a restricted area. The nodes in such a distributed system could be much more easily maintained since the burden of retraining or fine-tuning becomes partitioned into much smaller chunks. At the same time, such an approach would offer better resilience, since such a system has at worst only reduced functionality if a sub-unit fails, whilst a monolithic unit can only fail catastrophically.

This would also facilitate the application of RAG techniques to swap in locally specific knowledge (Lewis et al. 2020; Fan et al. 2024), and *cooperation with Digital Twins (DTs)*: DT instances typically represent models that are in correspondence to many kinds of physical subsystems in a city (such as road segments, buildings, or vehicles), which makes them tightly integrated into the "physical fabric" of a city (Ferré-Bigorra et al. 2022).

The flexible granularity of SLM arrays offers an ideal match for this granularity of DT systems, allowing each DT to become coupled with one or more dedicated SLMs from the array. In this coupling, the use of an SLM can extend well beyond just merely providing a natural language interface. Instead, the good tunability (or even trainability from scratch) of SLMs allows the interface to be *knowledgeable and specialized* to the task domain of the DT it is associated with (Fig. 7).

Such *knowledge-integrated* DT-SLM-pairs offer a new and very powerful building block to endow a city with a kind of virtual and distributed *"intelligent ecosystem"*. Each node of this ecosystem integrates subsymbolic control knowledge in the DT with natural language accessible symbolic and high-level knowledge in the associated SLM. This network can be hierarchical, with super-ordinate nodes supervising more local clusters. However, the selection of collaborating partners in such a large-scale system has been identified as a challenging problem (Rothermel et al. 2025). In the IoT domain, collaboration partners are often selected based on spatial proximity. This is in line with Tobler's observation (Tobler 1970) that "everything is related to everything else, but nearby things are more related than distant things". In Rothermel et al. (2025), we propose a proximity model that allows agents to identify collaborators based on spatial proximity. However, the concept of proximity is not limited to the spatial dimension and could therefore be extended to cognitive, institutional, social or organizational aspects.

Combining SLMs and Digital Twins and embedding them in a hierarchical network structured according to the spatial proximity structure of the city leads to an architecture which is very favorable for computational parsimony. While a more detailed analysis has to be postponed to a forthcoming paper, there exist already ample results how the utilization of SLMs can lead to significant computational gains. For instance, Irugalbandara et al. (2024) find in their study of 29 recent SLMs that most of them reach response qualities that are close to GPT- 4 level while incurring only a small fraction (1/5 or less) of the cost. The recent development of scalable matrix-multiplication free language models brings further reductions into reach. Zhu et al. (2024b) reports an FPGA implementation of a 1 billion parameter SLM employing this technique that consumes 13 W of power while generating text at humen reading speed.

The implementation of such an SLM-DT based "intelligent ecosystem" can provide a powerful infrastructure for planners to interact with the city, to monitor city states, or even to simulate the response of the city to unforeseen impacts. Its accessibility through natural language also enables its use to capture feedback or sentiments from citizens, to engage in distributed discussions about planning options, or to support forms of participatory

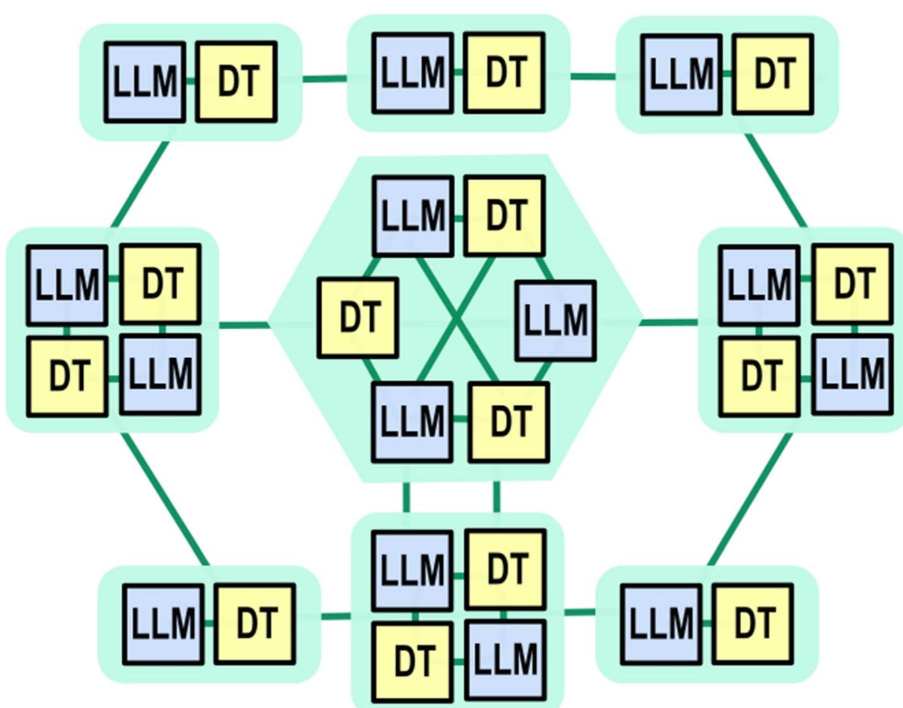

Fig. 7 Interconnected pairs of locally specialized Digital twins (DT) and language models (LLM) can provide powerful units for a distributed "Social AI System" within a city

planning that involve citizens via the interface. Unlike traditional forms of citizen participation, such a network might be able to evaluate the consistency of citizens' proposals, oversee compatibility with overarching long-term plans (e.g., such as the example of the Shanghai Masterplan (Leading Group Office 2017)), or evaluate feasibility with regard to resources and demands of involved players.

4.4 Ethical challenges

Urban planning creates long lasting impacts that affect the lives of many. This brings a large responsibility. At the same time, the multi-dimensional optimization expected from urban planners cannot be done without weighting and balancing a multitude of often conflicting optimization criteria and interests. This can lead into ethical dilemmata whose resolution will be affected by the tools that provide planners (and the public) with the models, simulations, perspectives and predictions for different planning scenarios and potential outcomes. A recent overview of major challenges is provided by Sanchez et al. (2024). They focus on the problems of bias, privacy, equity and inclusivity, accountability and transparency, danger of mis- and dis-information, and the necessity of human control and oversight to safeguard against limitations of AI, and they offer several recommendations in each of these areas. Kitchin (2016) takes a critical look at how the transition from small to big data influences our way how we conceive cities and their operational governance, with the danger of a "reduction of urban life to logic and calculative rules". He argues for a re-orientation in how cities are conceived and, likewise, emphasizes the importance of an ethics about privacy, harms, notice and consent, and recommends a pro-active stance of city managers. We would like to complement these broad analyses with three specific aspects that appear to us particularly focal to our work:

Offering intelligent agents roles of virtual citizens. Here, the created benefits come with the side-effect of potentially substituting real human action or feedback with that from a digital surrogate that may introduce (or even enhance) biases, compromise privacy (when the simulated behavior leaks individual information in the training set), or possibly endanger human participation and trust. This will require advances on the technology side in conjunction with clear legal frameworks to ensure that citizen models are based on training data that properly reflect local diversity patterns (e.g. gender, cultural, age) and are ethically sourced.

Pervasive urban AI with perfect recall. Endowing intelligent agents with more sophisticated memory systems (e.g., episodic memories) and connecting larger and larger parts of a city to sensors that provide continuous input to AI models can lead to a pervasive urban AI that has perfect recall of almost everything that has happened. We need to understand how we can appropriately limit such systems and endow them with a social AI that by its design will reliably evolve in harmony with human society. This might require to endow AI systems with some ethical understanding or memories with the ability to forget.

Virtual accessibility of "everything". To a large degree the significance of humans, experiences and things for us is related to the ways in which they are special to us. What will happen if sophisticated urban models reach a level of detail and realism that they can begin to compete with the real city and its citizens? How can we develop our ethics to provide guidance in future "hybrid cities" that not only mix cultures but also the real and the virtual in unprecedented seamlessness and variability of the virtual?

Thus, urban planning, AI and ethics need rich interactions that deserve careful attention to properly enhance cities as places that foster human potential and creativity.

5 Conclusions

Tracing back city models to their historical roots, we have shown their steady development from simple map-like descriptions to modern digital counterparts that combine many sophisticated methods to generate large scale renderings that can represent the physical substrate of cities to a high degree of realism. We then have set out to analyze what still is largely missing from these approaches: to adequately include the behavior of the citizens as the most important factor that an urban model must address. Linking this gap with the recent availability of very large-scale generative models with billions of parameters, particularly in the form of large language models (LLMs) and their recent multimodal variants, we have argued that we now are for the first time in a position to create rich citizen models that can capture a realistic level of complexity. We discussed strengths and weaknesses of these models, particularly their large resource footprints, the still existing challenges that come with adapting them in a fashion that is reliable and interpretable, and encouraging developments, particularly the advent of resource-parsimonious small language models (SLMs) that can approximate the capabilities of LLMs in much smaller footprints. We have then made a proposal how to bring these different strands of developments together to achieve a new level of urban modeling where models are interactive and can reach beyond traditional representations of a physical city populated by rather simple agents that are limited to an almost particle-like behavior due to their small number of parameters. Instead we have sketched a concept how we can exploit the capabilities of multimodal SLMs to reach into a modeling of the mental spaces that are contributed by the citizens and that we argue as being at the root of driving urban development. We have linked this to the concept of a “Social AI” that brings the citizens’ anchoring in a social world into its focus, along with important driving forces such as resources, desires, needs and sentiments. We finally came up with a sketch of an architecture that can provide the necessary substrate for such an endeavor. It achieves resource parsimony by employing a distributed network of nodes each of which combines SLMs and Digital Twins to effectively combine the holistic reasoning power and human-friendly interaction formats of the former with the hallucination-free reliability and inspectability of the latter. We argue that this approach can provide a solid technological base for realizing safe, manageable and resource parsimonious intelligent ecosystems for driving a new generation of urban models based on Social AI. We believe that this Social AI can foster not only the development of the physical spaces of the city but also the development of its social space which arises from the interaction of the many mental spaces of the citizens. While this certainly is still a distant vision, innovation from coupling urban models, Digital Twins and generative AI systems together can already now be seen to bring us ever closer to realising this exciting vision.

Appendix A ChatGPT- 4o outputs

Fig. A1: The ChatGPT-4o output on 4 Sept. 2024 given the image and prompt in Fig. 4.

The map you provided appears to depict the village of Pool-in-Wharfedale, located near the River Wharfe in the region of West Yorkshire, England. Here’s a detailed description of the geographical and built environment features visible on the map:

Geographical Features: 1. **River Wharfe**: A significant geographical feature on the map, the River Wharfe flows from the northwest to the east along the northern edge of the village. The river creates a natural boundary and likely influences the local landscape and activities.

2. **Leathley**: Across the River Wharfe, the area labeled as Leathley is indicated in North Yorkshire. The river serves as a boundary between West Yorkshire (where Pool-in-Wharfedale is located) and North Yorkshire.

3. **Green Spaces**: There are several green areas, particularly to the north and west of the village. These areas are likely open fields, woodlands, or agricultural lands, typical of the rural landscape in this region.

Built Environment Features: 1. **Main Roads**: - **A658**: A major road labeled A658 runs through the village from south to north, crossing the River Wharfe via a bridge. This road connects Pool-in-Wharfedale to surrounding towns and regions, making it a key transport route. - **A659**: This road intersects with the A658 in the southern part of the village, running from west to east, and continues towards the east as Arthington Lane.

2. **Residential Areas**: - The village is primarily composed of residential areas, with various roads such as Church Close, Mill Lane, and Manor Court serving the local community. These areas are characterized by clusters of houses and small streets.

3. **Community and Religious Buildings**: - **Churches**: There are two symbols of churches marked with crosses on the map. These are likely to be significant religious or community buildings. - **Pool-in-Wharfedale Church of England Voluntary Controlled Primary School**: Located towards the southeast of the village, this school is an important community institution.

4. **Recreational Areas**: - **Pool Cricket Club**: Situated to the northeast of the village, this cricket ground is a notable recreational facility for local residents. - The presence of several green spaces and the cricket club indicates recreational opportunities within the village.

5. **War Memorial**: A war memorial is marked in the central part of the village, indicating a place of remembrance and historical significance.

6. **Bridges**: The map shows a bridge where the A658 crosses the River Wharfe, which is a crucial piece of infrastructure connecting different parts of the region.

Surrounding Areas: - To the north, beyond the River Wharfe, the land appears to be more rural with less dense development, while the village itself is more built-up.

Overall, Pool-in-Wharfedale is depicted as a small village with essential amenities, recreational facilities, and residential areas, set against a backdrop of natural features like the River Wharfe and surrounding green spaces.

Fig. A2: ChatGPT-4o’s response on 4/9/2024 to the question “Please describe a route from the fishmonger to the school.”

To describe the route from the Fishmonger to the School on the provided map:

1. **Start at the Fishmonger:** - The Fishmonger is located on the right side of the map along the yellow road.

2. **Head left along the Yellow Road:** - Move left along the yellow road towards the center of the map. You will pass by the Restaurant, which is centrally located on this road.

3. **Continue past the Newsagent:** - Continue traveling left on the yellow road. You will pass the Newsagent, which is located to the left of the Restaurant.

4. **Reach the intersection with the Red Road:** - Continue until you reach the intersection where the yellow road crosses the red diagonal road.

5. **Turn left onto the Red Road:** - Turn left onto the red road.

6. **Proceed up the Red Road to the School:** - Continue traveling upward along the red road until you reach the School, which is located just slightly below the intersection of the red road and the yellow road.

This route will take you directly from the Fishmonger to the School by primarily using the yellow road and then switching to the red road near the School's location.

Fig. A3: The response of ChatGPT-4o on 4 Sept 2024 when presented with the sketch map and prompt displayed in Fig. 5.

The sketch map you've provided features a simple layout of a small town or village with various buildings and roads marked. Here's a detailed description that someone could use to recreate the map:

Roads: 1. **Red Road (Diagonal, Top-Left to Bottom-Right):** - Starts from the top left corner and ends at the bottom right corner. - Intersects with the green road and the yellow road.

2. **Green Road (Diagonal, Bottom-Left to Top-Right):** - Starts from the bottom left corner and ends at the top right corner. - Intersects with the red road and the yellow road.

3. **Yellow Road (Horizontal, Middle):** - Runs horizontally through the middle of the map. - Intersects with the red and green roads.

4. **Purple Road (Horizontal, Bottom):** - Runs horizontally near the bottom of the map. - Does not intersect with the red or green roads.

Buildings and Locations: - **Top-Left (near the Red Road):** - **School:** Positioned slightly below where the red road intersects with the yellow road.

- **Top-Right (along the Green Road):** - **Post Office:** Located just before the green road reaches the top-right corner.

- **Middle of the Map (on the Yellow Road):** - **Restaurant:** Centrally located on the yellow road, close to where it intersects with the green road.

- **Middle-Left (on the Yellow Road):** - **Newsagent:** Positioned to the left of the restaurant along the yellow road.

- **Middle-Right (on the Yellow Road):** - **Fishmonger:** Positioned to the right of the restaurant along the yellow road.

- **Bottom-Left (along the Purple Road):** - **Bakery:** Located at the far left on the purple road.

- **Bottom-Right (along the Purple Road):** - **Butcher:** Located at the far right on the purple road.

- **Middle (slightly off the Green Road):** - **Pub:** Positioned very close to the intersection of the green road and the yellow road, just below the yellow road.

- **Bottom-Left (along the Green Road):** - **Café:** Located where the green road intersects the purple road, but closer to the left side.

- **Bottom-Right (along the Red Road):** - **Charity Shop:** Located at the bottom right, close to where the red road meets the purple road.

Notes: - The buildings are depicted as small blue rectangles. - The roads are represented by colored lines: red, green, yellow, and purple. - The roads intersect at various points, creating a cross-like pattern with the red and green roads.

This detailed description should help someone recreate the map accurately.

Authors' contributions

H. R., O.H., K. R., A.G.C.: Conceptualization, writing, visualization, review, and editing, Z. W.: Conceptualization, review. All authors have read and agreed to the published version of the manuscript.

Funding

H. R., O.H., K. R., A.G.C. are partially funded by the Special Funds of Tongji University for the "Sino-German Cooperation 2.0 Strategy". A.G.C. also acknowledges support by the Fundamental Research priority area of The Alan Turing Institute.

Data availability

The detailed ChatGPT- 4o outputs upon prompts described in the paper are given in Appendix A. No other data are used in this research.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

H. R., K. R., A.G.C. declare no conflict of interests/competing interests. Z.W. is Editor-in-Chief of Frontiers of Urban and Rural Planning. O.H. is a member of the journal's editorial board.

Received: 15 October 2024 Revised: 12 March 2025 Accepted: 15 April 2025
Published online: 16 May 2025

References

Alavi, H. S., Churchill, E. F., Wiberg, M., Lalanne, D., Dalsgaard, P., Fatah gen Schieck, A., & Rogers, Y. (2019). Introduction to human-building interaction (hbi) interfacing hci with architecture and urban design. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *26*(2), 1-10.

Alsayed, S. S. (2024). Urban human needs: Conceptual framework to promoting urban city fulfills human desires. *Frontiers in Sustainable Cities, 6*, 1395980.
Axhausen, K. W., & Gärling, T. (1992). Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems. *Transport Reviews, 12*(4), 323–341.
Bachl, M. & Ferreira, D. C. (2019). City-GAN: Learning architectural styles using a custom conditional GAN architecture. arXiv preprint arXiv:1907.05280
Banerjee, S., Agarwal, A., & Singla, S. (2024). LLMs Will Always Hallucinate, and We Need to Live With This. https://doi.org/10.48550/arXiv.2409.05746
Barnsley, M. F. (2014). *Fractals everywhere*. Academic Press.
Batty, M., Longley, P., & Fotheringham, S. (1989). Urban growth and form: Scaling, fractal geometry, and diffusion-limited aggregation. *Environment and Planning A, 21*(11), 1447–1472.
Batty, M., & Longley, P. A. (1986). The fractal simulation of urban structure. *Environment and Planning A, 18*(9), 1143–1179.
Batty, M., & Xie, Y. (1996). Preliminary evidence for a theory of the fractal city. *Environment and Planning A, 28*(10), 1745–1762.
Baum, S. (2017). A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, 17–1. https://doi.org/10.2139/ssrn.3070741
Belussi, A., Migliorini, S., & Eldawy, A. (2022). Spatial embedding: a generic machine learning model for spatial query optimization. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'22. Association for Computing Machinery.
Bhandari, P., Anastasopoulos, A., & Pfoser, D. (2023). Are large language models geospatially knowledgeable? arXiv:2310.13002
Bhat, C. R. spsampsps Koppelman, F. S. (1999). *Activity-Based Modeling of Travel Demand* (pp. 35–61). Springer US.
Blender (2024). https://www.blender.org. Accessed 9 Dec 2024.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Chang, K.-H., Cheng, C.-Y., Luo, J., Murata, S., Nourbakhsh, M., & Tsuji, Y. (2021). Building-GAN: Graph-conditioned architectural volumetric design generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11956–11965).
Chen, J., Cohn, A. G., Liu, D., Wang, S., Ouyang, J., & Yu, Q. (2015). A survey of qualitative spatial representations. *The Knowledge Engineering Review, 30*(1), 106–136.
Chen, T., Li, Z., Xu, W., Zhu, Z., Li, D., Tian, L., Barsoum, E., Wang, P., & Cheng, J. (2024). TernaryLLM: Ternarized large language model. arXiv preprint arXiv:2406.07177
Chen, W. & Li, Z. (2024). Octopus v2: On-device language model for super agent. arXiv preprint arXiv:2404.01744
Chen, Y. (2009). Urban chaos and perplexing dynamics of urbanization. *Letters in Spatial and Resource Sciences, 2*, 85–95.
Chen, Y. (2021). Scaling, fractals and the spatial complexity of cities. In *Handbook on Cities and Complexity* (pp. 176–194). Edward Elgar Publishing.
Chen, Y., & Zhou, Y. (2006). Reinterpreting central place networks using ideas from fractals and self-organized criticality. *Environment and Planning B: Planning and Design, 33*(3), 345–364.
Cheng, Y. (2022). *Nonscientific*. A Study of Traditional Chinese Mapmaking. Springer Nature.
Clark, W. A., & Fossett, M. (2008). Understanding the social context of the schelling segregation model. *Proceedings of the National Academy of Sciences, 105*(11), 4109–4114.
Cohn, A. G., & Blackwell, R. E. (2024). Evaluating the ability of large language models to reason about cardinal directions. In *Proc. COSIT*. arXiv:2406.16528
Cohn, A. G. & Renz, J. (2008). Qualitative spatial representation and reasoning. In *Foundations of Artificial Intelligence* (vol. 3, pp. 551–596). Elsevier.
Couturier, A. (2024). Scenecity. https://www.cgchan.com. Accessed 9 Dec 2024.
Creswell, A. & Shanahan, M. (2022). Faithful reasoning using large language models. arXiv:2208.14271
Crooks, A., Heppenstall, A., Malleson, N., spsampsps Manley, E. (2021). Agent-based modeling and the city: A gallery of applications. *Urban Informatics*, 885–910.
Cugurullo, F. & Xu, Y. (2025). When AIs become oracles: generative artificial intelligence, anticipatory urban governance, and the future of cities. *Policy and Society*, *44*(1), 98-115
Das, D. & Kwek, B. (2024). AI and data-driven urbanism: The Singapore experience. *Digital Geography and Society*, 100104.
d'Avila Garcez, A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review, 56*(11), 12387–12406.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.
Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems, 34*, 8780–8794.
ESRI (2024). ArcGIS CityEngine - advanced 3D city design software. https://www.esri.comi/en-us/arcgis/products/arcgis-cityengine/overview. Accessed 9 Dec 2024.
Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6491–6501).
Fang, Z., Jin, Y., & Yang, T. (2022). Incorporating planning intelligence into deep learning: A planning support tool for street network design. *Journal of Urban Technology, 29*(2), 99–114.
Ferré-Bigorra, J., Casals, M., & Gangolells, M. (2022). The adoption of urban digital twins. *Cities, 131*, 103905.
Floridi, L. (2020). Artificial Intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy & Technology, 33*(4), 541–546.
Forrester, J. W. (1970). Urban dynamics. *Industrial Management Review (pre-1986)*, 11(3), 67.
Fotheringham, A. S., Batty, M., & Longley, P. A. (1989). Diffusion-limited aggregation and the fractal nature of urban growth. *Papers in Regional Science, 67*(1), 55–69.
Gibała-Kapecka, B. & Kapecki, T. (2016). The real city-the mental city. In *Back to the Sense of the City: International Monograph Book* (pp. 1402–1414). Centre de Política de Sòl i Valoracions.
González-Méndez, M., Olaya, C., Fasolino, I., Grimaldi, M., & Obregón, N. (2021). Agent-based modeling for urban development planning based on human needs. conceptual basis and model formulation. *Land Use Policy*, 101, 105110.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (NeurIPS 2014) (pp. 2672–2680).
Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2020). A review on generative adversarial networks: Algorithms, theory, and applications. arXiv preprint arXiv:2001.06937
Halsey, T. C. (2000). Diffusion-limited aggregation: A model for pattern formation. *Physics Today, 53*(11), 36–41.
Helbing, D. (2001). Traffic and related self-driven many-particle systems. *Reviews of Modern Physics, 73*(4), 1067.
Herzog, O., Jarke, M., & Zhiqiang, S. W. (2023). Cooperating and competing digital twins for industrie 4.0 in urban planning contexts. *Sci*, 5(4)(1), 44.
Hillier, B., Leaman, A., Stansall, P., & Bedford, M. (1976). Space syntax. *Environment and Planning B: Planning and Design, 3*, 147–185.
Hong, G., Gema, A. P., Saxena, R., Du, X., Nie, P., Zhao, Y., Perez-Beltrachini, L., Ryabinin, M., He, X., Fourrier, C., & Minervini, P. (2024). The Hallucinations Leaderboard – An Open Effort to Measure Hallucinations in Large Language Models. arXiv:2404.05904
Hong, Y., Hwang, U., Yoo, J., & Yoon, S. (2019). How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR), 52*(1), 1–43.
Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685
Huang, J. & Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. arXiv:2212.10403
Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232

Huang, Q., Parker, D. C., Filatova, T., & Sun, S. (2014). A review of urban residential choice models using agent-based modeling. *Environment and Planning B: Planning and Design, 41*(4), 661–689.
Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., Tang, L., Kang, Y., & Mars, J. (2024). Scaling down to scale up: A cost-benefit analysis of replacing openai's llm with open source slms in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 280–291). IEEE.
Isinkaralar, O., & Varol, C. (2023). A cellular automata-based approach for spatio-temporal modeling of the city center as a complex system: The case of Kastamonu. *Türkiye. Cities, 132*, 104073.
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1125–1134).
Jahanmiri, F., & Parker, D. C. (2022). An overview of fractal geometry applied to urban planning. *Land, 11*(4), 475.
Jensen, P. (2022). Introducing simple models of social systems. *American Journal of Physics, 90*(6), 462–468.
Jiang, F., Ma, J., Webster, C. J., Li, X., & Gan, V. J. (2023). Building layout generation using site-embedded GAN model. *Automation in Construction, 151*, 104888.
Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications, 207*, 117921.
Kastner, P., & Dogan, T. (2023). A GAN-based surrogate model for instantaneous urban wind flow prediction. *Building and Environment, 242*, 110384.
Kelly, G., & McCabe, H. (2006). A survey of procedural techniques for city generation. *The ITB Journal, 7*(2), 5.
Kim, J.-S., Kavak, H., & Crooks, A. (2018). Procedural city generation beyond game development. *SIGSPATIAL Special, 10*(2), 34–41.
Kitchin, R. (2016). The ethics of smart cities and urban science. *Philosophical transactions of the royal society A: Mathematical, physical and engineering sciences, 374*(2083), 20160115.
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (NeurIPS 2012) (pp. 1097–1105).
Leading Group Office (2017). Shanghai Master Plan 2017–2035.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
Li, B., Mellou, K., Zhang, B., Pathuri, J., & Menache, I. (2023a). Large language models for supply chain optimization. arXiv:2307.03875
Li, F., Hogg, D. C., & Cohn, A. G. (2024). Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the StepGame benchmark. *In Proceedings of the AAAI Conference on Artificial Intelligence, 38*, 18500–18507.
Li, X., Lu, W., Peng, Z., Zhang, Y., & Huang, J. (2024b). Generative design of walkable urban cool spots using a novel heuristic GAN× GAN approach. *Building and Environment*, 112027.
Li, X., Zhang, Y., Li, C., Wang, T., & Xi, C. (2022a). Application of digital twin in handling and transportation of hazardous chemicals. *Applied Sciences*, 12(24). https://doi.org/10.3390/app122412746
Li, Z., Kim, J., Chiang, Y.-Y., & Chen, M. (2022b). SpaBERT: A pretrained language model from geographic data for geo-entity representation. arXiv:2210.12213
Li, Z. & Ning, H. (2023). Autonomous GIS: the next-generation AI-powered GIS. arXiv:2305.06453
Li, Z., Zhou, W., Chiang, Y.-Y., & Chen, M. (2023b). GeoLM: Empowering Language Models for Geospatially Grounded Language Understanding. arXiv:2310.14478
Ling, C., Niu, X., Yang, J., Zhou, J., & Yang, T. (2024). Unravelling heterogeneity and dynamics of commuting efficiency: Industry-level insights into evolving efficiency gaps based on a disaggregated excess-commuting framework. *Journal of Transport Geography, 115*, 103820.
Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., & Lao, N. (2023). On the opportunities and challenges of foundation models for geospatial artificial intelligence. arXiv:2304.06798
Mandelbrot, B. B. (1983). *The fractal geometry of nature* (Revised and enlarged ed.). W. H. Freeman and Co
Mansourian, A. & Oucheikh, R. (2024). ChatGeoAI: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10), 348.
Mehaffy, M. W. & Elmlund, P. (2023). Smart cities: Missing the stigmergy? In *Knowledge Management for Regional Policymaking* (pp. 149–161). Springer.
Midjourney (2024). https://www.midjourney.com/. Accessed 9 Dec 2024.
Mihai, S., Yaqoob, M., Hung, D. V., Davis, W., Towakel, P., Raza, M., Karamanoglu, M., Barn, B., Shetve, D., Prasad, R. V., et al. (2022). Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Communications Surveys & Tutorials, 24*(4), 2255–2291.
Molinero, C., & Thurner, S. (2021). How the geometry of cities determines urban scaling laws. *Journal of the Royal Society interface, 18*(176), 20200705.
Moody, H. T. (1970). Urban dynamics: A review of Forrester's model of an urban area. *Economic Geography, 46*(4), 620–626.
Nauata, N., Chang, K.-H., Cheng, C.-Y., Mori, G., & Furukawa, Y. (2020). House-GAN: Relational generative adversarial networks for graph-constrained house layout generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (pp. 162–177). Springer
Pal, D., Triyason, T., & Padungweang, P. (2018). Big data in smart-cities: Current research and challenges. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 6*(4), 351–360.
Parish, Y. I. & Müller, P. (2001). Procedural modeling of cities. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (pp. 301–308).
Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442
Patil, D., Rane, N., Rane, J., & Paramesha, M. (2024). Artificial intelligence and generative AI, such as ChatGPT, in transportation: Applications, technologies, challenges, and ethical considerations. *Trustworthy Artificial Intelligence in Industry and Society*, 185–232.
Prusinkiewicz, P. (1986). Graphical applications of l-systems. *In Proceedings of Graphics Interface, 86*, 247–253.
Qi, J., Li, Z., & Tanin, E. (2023). MaaSDB: Spatial databases in the era of large language models (vision paper). In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23. ACM.
Quan, S. J. (2022). Urban-GAN: An artificial intelligence-aided computation system for plural urban design. *Environment and Planning B: Urban Analytics and City Science, 49*(9), 2500–2515.
Redding, S. J., & Rossi-Hansberg, E. (2017). Quantitative spatial economics. *Annual Review of Economics, 9*, 21–58.
Rothermel, K., Herzog, O., & Wu, Z. (2025). Proximity-based service discovery for distributed digital twin systems. https://doi.org/10.1007/s43926-025-00103-x
Roumpani, F. (2022). Procedural cities as active simulators for planning. *Urban Planning, 7*(2), 321–329.
Ruiz, A., Melnik, A., Wang, D., & Ritter, H. (2024). Lane segmentation refinement with diffusion models. arXiv:2405.00620
Sanchez, T. W., Brenman, M., & Ye, X. (2025). The ethical concerns of artificial intelligence in urban planning. *Journal of the American Planning Association*, 1–14.
Sanchez-Sepulveda, M. V., Navarro, J., Amo-Filva, D., Fonseca, D., Antúnez-Anea, F., & Barranco-Albalat, A. (2024). A data-driven approach to enhance urban infrastructure for sustainable mobility and improved quality of life in highly populated cities. case study: Barcelona. *Frontiers in Built Environment*, 10.
Shen, G. (2002). Fractal dimension and fractal growth of urbanized areas. *International Journal of Geographical Information Science, 16*(5), 419–437.
Siegenfeld, A. F., & Bar-Yam, Y. (2020). An introduction to complex systems science and its applications. *Complexity, 2020*(1), 6105872.
Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502
Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies, 61*, 49–62.

Sun, Z., Zheng, J., & Hu, H. (2012). Fractal pattern in spatial structure of urban road networks. *International Journal of Modern Physics B, 26*(30), 1250172.

Theobald, U. (2010). Chinaknowledge.de - an encyclopaedia on Chinese history and literature. https://www.chinaknowledge.de/Literature/Science/jingdingjiankangzhi.html. Accessed 9 Nov 2024.

Tobler, W. (1979). Cellular geography. *Philosophy in Geography*.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, Supplement: Proceedings, 46*, 234–240.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv:2401.01313

Unlu, E. (2023). Chatmap : Large language model interaction with cartographic data. arXiv:2310.01429

Van Nes, A. & Yamu, C. (2021). *Introduction to space syntax in urban studies*. Springer Nature.

Vanegas, C., Aliaga, D., Wonka, P., Müller, P., Waddell, P., & Watson, B. (2009). Modelling the appearance and behaviour of urban spaces. In *Computer graphics forum* (pp. 25–42). Oxford: Blackwell Publishing Ltd.

Wang, D., Fu, Y., Liu, K., Chen, F., Wang, P., & Lu, C.-T. (2023). Automated urban planning for reimagining city configuration via adversarial learning: Quantification, generation, and evaluation. *ACM Transactions on Spatial Algorithms and Systems, 9*(1), 1–24.

Wang, J., Jiang, R., Yang, C., Wu, Z., Onizuka, M., Shibasaki, R., Koshizuka, N., & Xiao, C. (2024a). Large language models as urban residents: An LLM agent framework for personal mobility generation. arXiv preprint arXiv:2402.14744

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024b). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345.

Wang, S., Zeng, W., Chen, X., Ye, Y., Qiao, Y., & Fu, C.-W. (2021). Actfloor-GAN: Activity-guided adversarial networks for human-centric floorplan design. *IEEE Transactions on Visualization and Computer Graphics, 29*(3), 1610–1624.

Wang, X., Ling, X., Zhang, T., Li, X., Wang, S., Li, Z., Zhang, L., & Gong, P. (2023b). Optimizing and fine-tuning large language model for urban renewal. arXiv preprint arXiv:2311.15490

White, R., & Engelen, G. (2000). High-resolution integrated modelling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems, 24*(5), 383–400.

Witten, T. A., & Sander, L. M. (1983). *Diffusion-limited aggregation. Physical Review B, 27*(9), 5686.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). Autogen: Enabling Next-Gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155

Xie, H., Chen, Z., Hong, F., & Liu, Z. (2024). Citydreamer: Compositional generative model of unbounded 3d cities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9666–9675).

Xu, C., Liu, M., Zhang, C., An, S., Yu, W., & Chen, J. M. (2007). The spatiotemporal dynamics of rapid urban growth in the Nanjing metropolitan region of China. *Landscape Ecology, 22*, 925–937.

Xu, F., Li, Y., & Song, C. (2021). Emergence of urban growth patterns from human movements. *Nat Comput Sci*, *1*, 791–800. https://doi.org/10.1038/s43588-021-00160-6

Xu, P., Broniatowski, D. A., & Dredze, M. (2024). Twitter social mobility data reveal demographic variations in social distancing practices during the covid-19 pandemic. *Scientific Reports, 14*(1), 1165.

Xu, Z., Jain, S., & Kankanhalli, M. (2024b). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys, 56*(4), 1–39.

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F., & Pozdnoukhov, A. (2018). A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems, 19*(6), 1682–1696.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2022). Deep learning on traffic prediction: Methods, analysis, and future directions. *IEEE Transactions on Intelligent Transportation Systems, 23*(6), 4927–4943.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). MM-LLMs: Recent advances in multimodal large language models. arXiv:2401.13601

Zhang, H., Lan, T., & Li, Z. (2022). Fractal evolution of urban street networks in form and structure: A case study of Hong Kong. *International Journal of Geographical Information Science, 36*(6), 1100–1118.

Zhang, J. (2021). Agent-based optimizing match between passenger demand and service supply for urban rail transit network with NetLogo. *IEEE Access, 9*, 32064–32080.

Zhang, X., Shehata, A., Benes, B., & Aliaga, D. (2020). Automatic deep inference of procedural cities from global-scale spatial data. *ACM Transactions on Spatial Algorithms and Systems (TSAS), 7*(2), 1–28.

Zhou, Z., Li, L., Chen, X., & Li, A. (2023). Mini-giants: "small" language models and open source win-win. arXiv preprint arXiv:2307.08189

Zhu, H., Zhang, W., Huang, N., Li, B., Niu, L., Fan, Z., Lun, T., Tao, Y., Su, J., Gong, Z., Fang, C., & Liu, X. (2024a). PlanGPT: Enhancing urban planning with tailored language model and efficient retrieval. arXiv:2402.19273

Zhu, R.-J., Zhang, Y., Sifferman, E., Sheaves, T., Wang, Y., Richmond, D., Zhou, P., & Eshraghian, J. K. (2024b). Scalable matmul-free language modeling. arXiv preprint arXiv:2406.02528

Publisher's Note