Urban Planning
2026 • Volume 11 • Article 12126
https://doi.org/10.17645/up.12126

ARTICLE

Open Access Journal

Creating Synthetic Road Networks Using Large Language Models: An Iterative Calibration Framework

Adam Kasiński and Przemysław Szufel

Decision Analysis and Support Unit, SGH Warsaw School of Economics, Poland

Correspondence: Adam Kasiński (ak116751@doktorant.sgh.waw.pl)

Submitted: 8 February 2026 Accepted: 27 April 2026 Published: 27 July 2026

Issue: This article is part of the issue "The Urban Canvas: Spatial Computing in Planning, Analysing, and Representing the Cityscape" edited by Valerio Signorelli (University College London), Andy Hudson-Smith (University College London), and in memory of Heejung Kwon (Yonsei University), fully open access at https://doi.org/10.17645/up.i484

Abstract

Digital twins are used to represent and analyse urban environments, providing a basis for simulation, scenario testing, and communication among stakeholders. While effective, such models are typically developed as unique, data-intensive representations of specific cities. Synthetic cities extend this paradigm by enabling the systematic generation of entire families of digital twins. They provide controlled, repeatable, and easily calibratable environments, making it possible to conduct analyses that are more general, reproducible, and independent of costly or sensitive field data. Generating synthetic cities addresses several key needs in urban research. It supports the design and evaluation of regulatory scenarios, facilitates climate adaptation analyses, and enriches participatory design by providing semantically meaningful models for discussion. By decoupling experimentation from real-world data constraints, synthetic cities provide a robust framework for testing alternative urban forms and policies while preserving privacy and reducing costs. This article introduces a novel method for generating synthetic cities using large language models (LLMs). Specifically, we propose an iterative LLM-guided framework in which the model directly generates high-level synthetic road structures and calibrates them through tool-based structural feedback. Unlike procedural generators and neural generative models, which usually require manual parameter tuning and offer limited ways to enforce high-level semantic constraints, our LLM-based framework translates natural-language requirements into quantitative specifications and validates generated networks against target metrics. This supports the creation of adaptable, semantically meaningful synthetic cities suitable for simulation and scenario analysis.

Keywords

digital twins; large language models; synthetic cities; urban modelling

© 2026 by the author(s), licensed under a Creative Commons Attribution 4.0 International License (CC BY).

1. Introduction

In recent years, spatial planning decisions have increasingly relied on simulation methods such as agent-based models (Barthelemy, 2019). Bonabeau (2002) points out that approaches grounded in microfoundations are especially valuable because they capture the complex behaviours of individual agents that cannot be properly represented in an aggregated model. However, these simulations require realistic environments in which agents can operate, and digital models—or "synthetic cities"—have recently become popular for this purpose, building on earlier work on virtual urban representations for visualisation, simulation, and scenario exploration (Batty & Hudson-Smith, 2001; Bilal et al., 2025). These synthetic environments help policymakers forecast the outcomes of their decisions (Zheng et al., 2023). At the same time, agent-based models are prone to artefacts that may be difficult to trace (Galán et al., 2009). Therefore, when testing new urban policies in a virtual city model, more robust conclusions may be obtained by evaluating them on a population of cities with similar structural properties rather than on a single city instance. Kim et al. (2018) argue that synthetic cities are particularly useful when they preserve the key structural features of real-world agglomerations—such as their transportation networks—while abstracting from irrelevant city-specific details, thereby providing standardised testbeds for policy evaluation. Such structures have been applied in contexts ranging from autonomous vehicle trials (Kim et al., 2018) and resilience analyses (Therias & Rafiee, 2023) to energy-demand forecasting (Roth et al., 2020).

Beyond enabling cross-regional comparability, synthetic cities also provide reproducibility for controlled experiments (Barbie & Hasselbring, 2024), cost-effective scenario testing without field deployments (Almirall et al., 2022), and privacy-preserving abstractions of sensitive urban data (Papyshev & Yarime, 2021). Xu et al. (2024) further point out that synthetic cities enable benchmarking of algorithms—from traffic-flow optimisers to emergency-evacuation planners—under identical conditions and facilitate rapid "what-if" analyses of land-use changes or infrastructure investments.

Existing approaches to synthetic city generation can be grouped into six families according to how the generation process is controlled: procedural (rule-based), graph-based (roads represented as abstract networks), tensor-field-based (street orientations encoded as spatial fields), simulation-based (agent- or process-based simulation driving growth), example-based (stitching together real-world fragments), and learning-based (neural generative models). A detailed review of these families is provided in Section 2.

To compare these approaches, we focus on four capabilities that are particularly important for using synthetic cities as reusable experimental testbeds:

1. Calibration: Adjusting the generator so that selected output metrics match a reference city;
2. Conditional attribute controllability: Explicitly constraining the generated city to satisfy user-specified quantitative or semantic attributes;
3. Zero-shot customisation: Supporting novel, user-specified design requirements without retraining or fine-tuning the generator;
4. Semantic reasoning: Interpreting high-level, natural-language urban concepts and translating them into consistent structural constraints.

As shown in Section 2 and summarised in Table 1, existing method families typically support only a subset of these capabilities. In particular, no single non-large-language-model (LLM) approach combines calibration and conditional control while also allowing zero-shot customisation, which motivates the contribution of this article.

The proposed approach offers two major benefits for digital twin developers. First, it decouples experimentation from real-world data, enabling easier construction of spatial simulation models. Second, it supports generating families of cities with desired spatial properties, which enables the development of more robust simulation models of urban environments.

The article is organised as follows. After this introduction, Section 2 presents the relevant literature. Section 3 introduces the proposed model, and Section 4 describes the experimental setup and evaluation procedures. Section 4.1 presents the experimental design, Section 4.2 reports the empirical results, and Section 4.3 discusses their implications. Finally, Section 5 concludes.

2. Literature Background

The goal of this section is to review the existing literature and motivate our contribution. The relevant work can be divided into three groups: synthetic city generation methods and their properties (Section 2.1), network comparison metrics used to assess the quality of a generated synthetic city (Section 2.2), and LLMs in the context of synthetic data modelling (Section 2.3). After reviewing these strands in Section 2.4, we present our contribution.

2.1. Synthetic City Generation Methods

Synthetic city generation methods can be grouped into six broad categories: procedural, graph-based, tensor-field-based, simulation-based, example-based, and learning-based. Each category is discussed in detail below.

Procedural methods generate cities through predefined rules. Parish and Müller (2001) proposed an extensible framework based on an enhanced L-system, in which roads are generated from production rules derived from global objectives and local constraints, after which the urban area is subdivided into plots and populated with building geometries. This design also allows additional city layers, such as points of interest, to be aligned with the road network. Weber et al. (2009) extended this approach to support more irregular and organically evolving layouts, while Beneš et al. (2014) introduced a dynamic variant in which trade, water transportation, and traffic simulation drive road-network growth over time. Overall, procedural methods are flexible and can generate complete, even dynamically evolving urban structures, but their outputs depend strongly on the predefined rules, parameters, and constraints.

Graph-based approaches represent roads as abstract networks and generate synthetic cities by reproducing structural properties of real-world road systems, such as hierarchy, sparsity, connectivity, and routing behaviour. In this line of work, Bauer et al. (2010) compared a Voronoi-based generator (Fortune, 1986) and the Abraham et al. road-network generator (Abraham et al., 2010) with OpenStreetMap-derived networks using structural and routing-based measures, including contraction hierarchies (Geisberger et al., 2008).

Both generators outperformed simple grid and unit-disk baselines, while the Voronoi-based approach more closely matched the hierarchy, sparsity, and routing performance of actual road systems. Lüscher and Weibel (2013) extended graph-based city modelling by incorporating semantic and functional data, such as topographic databases, points of interest, and user annotations. Overall, graph-based methods reproduce key structural properties of real-world road networks, but they may still generate unrealistic characteristics, such as excessive density or overly high node degrees.

Tensor-field approaches encode local street orientations in a spatial field and generate road networks by tracing streets along tensor-derived direction fields. G. Chen et al. (2008) introduced this approach by encoding local orientations through tensor eigenvectors and tracing streamlines to form hierarchical street networks. Sun and Dogan (2023) later extended the framework by allowing each tensor cell to store not only directional information but also land-use priorities, population density, and other non-geometric factors. In this way, tensor-field methods generate coherent directional patterns and can incorporate additional spatial information, although the resulting network depends strongly on what is encoded in the tensor field.

Zhang and Shi (2022) integrated a tensor-field framework with a simulation-based approach, illustrating the broader logic of simulation-based methods. In this family, road networks emerge through dynamic processes, making these approaches well-suited to capturing evolving and naturalistic urban structures. Their main advantage is the ability to represent hierarchy and local adaptation over time. In related work, Song and Whitehead (2019) proposed an agent-based method in which roads emerge through agents' interactions with resource nodes. Their main limitation is that the resulting form depends strongly on the behavioural rules, thresholds, and environmental factors specified in the simulation.

Example-based methods construct new layouts by copying and stitching fragments derived from real-world cities. Yu and Steed (2012) modelled existing urban structures as graphs and synthesised new nodes and edges by matching unfinished nodes with similar samples using a neighbourhood similarity measure, thereby supporting both entirely new structures and incremental network growth. Nishida et al. (2016) combined example-based and procedural techniques by fitting distinctive subgraphs from real-world cities into the target network and using a procedural fallback, guided by statistical features such as average road length and curvature, when no suitable patch was available. Overall, example-based methods preserve characteristic patterns from real-world cities, but the generated structure depends strongly on the quality and representativeness of the source examples.

Learning-based methods generate synthetic road networks by training neural models on examples of real urban layouts. Their main advantage is the ability to learn complex structural regularities directly from data, including patterns related to street density, hierarchy, connectivity, and block structure. Existing approaches include image-based models trained on rasterised road maps, graph-based models that operate directly on street-network representations, and sequential generative models that construct networks step by step. For example, Hartmann et al. (2017) used a generative adversarial network on binary rasters of OpenStreetMap road networks, Neira and Murcio (2022) proposed a graph-based approach combining a transformer with a variational graph auto-encoder, and Chu et al. (2019) introduced a sequential model, called "neural turtle graphics," for vector road graphs. Overall, these methods can capture rich structural patterns, but they typically require large training datasets and offer limited flexibility for imposing new constraints without retraining.

2.2. Network Comparison Metrics

Urban transport networks can be compared using methods that emphasise functional performance, structural topology, or latent graph representations. Existing work can be grouped into four broad families: (a) transport and routing metrics that quantify functional similarity; (b) graph edit and optimal-transport methods; (c) topological data analysis that captures higher-order connectivity and shape; and (d) graph kernels and neural approaches that embed graphs into feature spaces for statistical comparison. The following review summarises representative studies in each category.

Transport metrics compare urban networks by functional similarity rather than purely geometric or topological form. Common measures include density, degree distribution, shortest-path distance distributions, average road-segment length, intersection density, and circuity, which together capture connectivity, navigability, and route directness (Bauer et al., 2010; Neira & Murcio, 2022). The average path length similarity (APLS) metric extends this perspective by measuring mean relative deviations in shortest-path lengths between corresponding nodes (Chu et al., 2019).

Graph edit distance quantifies structural dissimilarity via the "cost" of transforming one transport network into another. For example, Li et al. (2014) adopt an optimal-transport perspective: They represent each city's metro system as a graph and measure pairwise differences with the Wasserstein distance, capturing both spatial arrangement and topology in a geometry-aware comparison.

A further line of work compares the topological shape of transport networks using topological data analysis (TDA), most commonly persistent homology. Feng and Porter (2020) used persistence diagrams to quantify connectivity and loops and to classify cities into broad morphological groups that differ in regularity and spatial organisation. Relatedly, Corcoran and Jones (2022) introduced persistence density functions to characterise connectivity across street classes and compared these distributions using the Wasserstein distance.

Another strand compares urban networks in an embedding space using graph kernels and graph neural networks (GNNs). Graph kernels represent each graph through features capturing local or global connectivity. For example, the shortest-path kernel compares all-pairs shortest-path lengths to measure similarity in overall distance structure, whereas the graphlet kernel counts small induced subgraphs, such as triangles and 4-cycles, to capture local motifs (Borgwardt & Kriegel, 2005; Shervashidze et al., 2009).

2.3. LLMs

Due to the novelty of this area, only a few studies have applied LLMs to synthetic city generation. Long et al. (2024) point out that LLMs have been used successfully for generating synthetic data. Rasal and Boddhu (2024) introduced NavGPT, a multimodal LLM that ingests aerial imagery and directly outputs precise road-network geometries. J. Chen et al. (2026) proposed network generation artificial intelligence (NGAI), an LLM-driven system that orchestrates specialised external plugins to produce fully routable road networks from either natural-language or image prompts. These studies demonstrate the potential of LLMs in road-network generation, but they do so either through multimodal input processing or by using the model primarily as an orchestration layer for external generation modules. By contrast, our article proposes an iterative LLM-guided framework in which the model itself directly generates and calibrates high-level synthetic road structures using tool-based structural feedback. The novelty of our approach therefore lies in using the LLM itself as the core generator within an iterative calibration loop, rather than primarily as an orchestration layer for specialised external modules.

2.4. Contribution

The desired capabilities of synthetic city generators identified in the literature include the ability to calibrate against real-world data, conditional attribute control, zero-shot customisation, and the inclusion of semantic reasoning in the generation process. Calibration against real-world data involves adjusting the parameters of synthetic city models so that generated outcomes align closely with real-world urban metrics, such as road density, street connectivity, and traffic patterns (van Vliet, 2013). Conditional attribute control allows users to explicitly specify characteristics or constraints for city generation. For instance, users might define transportation network requirements, enabling highly customised synthetic urban network generation tailored to specific analytical or planning objectives (Gan et al., 2024). Zero-shot customisation describes the ability of generation systems to respond to novel design requirements without prior task-specific training or parameter tuning. In the present article, this capability is studied in a reference-conditioned setting, where the framework generates calibrated variants of a reference urban structure in response to new prompts rather than unconstrained cities from scratch. This feature is particularly valuable for exploring new scenarios quickly and efficiently, significantly extending the versatility and applicability of synthetic city generation (Rasal & Boddhu, 2024). At present, conventional methods do not support this capability. Finally, semantic reasoning involves the capacity to interpret and incorporate conceptual urban knowledge within the generation process. This might include identifying appropriate zones for residential, commercial, or industrial development, or placing amenities and infrastructure in a way that reflects high-level urban planning principles (Webb et al., 2023).

Existing families of methods (procedural, graph-based, tensor-field, simulation, example-based, deep-learning) each miss one or more important capabilities—such as zero-shot customisation, or true semantic reasoning—while only LLMs theoretically cover them all (Table 1). Moreover, prior systems either depend on specialised external modules or target only narrow facets (e.g., roads), and cannot spontaneously capture higher-order phenomena like emergent city centres without explicit rules. The literature lacks an end-to-end, LLM-guided approach for calibrated, vector-based synthetic city generation that affords conditional attribute control, zero-shot modifications, and semantic understanding of urban structure (Brown et al., 2020; Webb et al., 2023). Our article fills this gap by proposing exactly such an LLM-centric framework.

Table 1. Synthetic city generation methods comparison by selected features.

Feature	Procedural	Graph-based	Tensor Fields	Simulation-based	Example-based	Deep learning-based	LLM
Is calibrated	X	X	✓	X	✓	✓	✓
Conditional attribute controllability	X	X	X	X	X	✓	✓
Zero-shot customisation	X	X	X	X	X	X	✓
Semantic reasoning	X	X	X	X	X	X	✓

3. Synthetic City Generation

The proposed framework focuses on high-level road-network structure within synthetic cities. This section presents a novel LLM-backed method for synthetic road-network generation. It has two parts. First, we formalise the road-network model and introduce a distance metric for comparing synthetic road networks. Next, we describe the proposed LLM-based generation framework, which produces synthetic road networks that satisfy both qualitative user requirements and quantitative calibration constraints.

3.1. City Modelling

3.1.1. Road Network

We assume that the road network is represented in the form of a weighted, undirected (simple) graph $G = (V, E, w)$, where V is a finite set of vertices, $E \subseteq \{\{u, v\} \subseteq V : u \neq v\}$ is the set of edges of the graph (represented as unordered pairs of vertices), and $w : E \to \mathbb{R}_+$ assigns a positive weight to each edge.

A synthetic city is represented by $\mathcal{C} = (G, \phi, \kappa)$, with the underlying graph $G = (V, E, w)$ representing the road network. Vertices correspond to junctions, edges correspond to road segments, $\phi : V \to \mathbb{R}^2$ assigns planar (ENU) coordinates to vertices, and $\kappa : E \to \mathcal{K}$ assigns a road class to each edge where $\mathcal{K} = \{1, \dots, K\}$, $K \in \mathbb{N}$ is a set of road classes. We denote by $\mathfrak{C}$ the set of all synthetic cities of this form.

3.1.2. City Similarity Distance

Let $\mathcal{C}_1 = (G_1, \phi_1, \kappa_1)$ and $\mathcal{C}_2 = (G_2, \phi_2, \kappa_2)$ be two synthetic cities. A similarity distance between cities is a function:

$$d : \mathfrak{C} \times \mathfrak{C} \to \mathbb{R}_{\geq 0} \tag{1}$$

We define d as an edge-length distribution discrepancy. $d\left(\mathcal{C}_1, \mathcal{C}_2\right)$ measures the discrepancy between the empirical distributions of road-segment lengths in the two cities. Let L_1 and L_2 denote the random variables "road-segment length" in cities $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, and let F_1 and F_2 be their cumulative distribution functions (see Togninalli et al., 2019). We define:

$$d\left(\mathcal{C}_1, \mathcal{C}_2\right) = \int_0^1 \left| F_1^{-1}(q) - F_2^{-1}(q) \right| \, dq \tag{2}$$

Where $F_i^{-1}(q) = \inf\{x : F_i(x) \geq q\}$ is the quantile function and $q \in [0, 1]$.

3.1.3. Road Network Compression

Road networks extracted from real-world cities often contain many degree-2 nodes that do not represent actual junctions or structural features, but instead encode road curvature. To simplify the graph for topological analysis, we apply a node-compression operation that removes these degree-2 nodes while preserving the network's overall connectivity and distances.

For the road network $G = (V, E, w)$ we define a node compression operation. Let $\deg(v) = |\{u \in V : \{v, u\} \in E\}|$ be the degree of a vertex $v \in V$. Consider a path $P_{\deg=2}$ in graph G, $P_{\deg=2} = (v_0, v_1, \dots, v_k)$, $(v_0 \neq v_1 \neq \dots \neq v_k)$,

$k \geq 2$, $\{v_i, v_{i+1}\} \in E$ such that the path's internal vertices all have degree 2 (i.e., $\deg(v_1) = \cdots = \deg(v_{k-1}) = 2$) and the first and last vertex have a different degree than 2 (i.e., $\deg(v_0) \neq 2 \wedge \deg(v_k) \neq 2$). We define *path compression* as deleting the vertices $v_1, \ldots, v_{k-1}$ from G and replacing the chain $(v_0, v_1, \ldots, v_k)$ with a single edge $\{v_0, v_k\}$. The weight of the new edge is set to the length of the compressed path, i.e.:

$$w'(\{v_0, v_k\}) = \sum_{i=0}^{k-1} w(\{v_i, v_{i+1}\})$$

If an edge $\{v_0, v_k\}$ already exists in G, then in the compressed graph we keep a single edge between v_0 and v_k with the minimum weight (distance). Additionally, if a connected component of G is a cycle whose all vertices have degree 2, then we delete this component entirely during compression. By the *compressed graph* $G' = (V', E', w')$ we mean the graph obtained after compressing all $P_{\deg=2}$ paths. Note that the compressed graph contains no vertices of degree 2.

3.2. Method

The method aims to generate a synthetic road network that satisfies both qualitative user requirements and quantitative similarity constraints with respect to a reference city. The workflow comprises two stages and is illustrated in Figure 1. The first stage produces a planar, high-level road network using an LLM-guided iterative loop with external validation tools. The second stage refines the accepted high-level structure by generating lower-level streets within planar regions using a Voronoi-based procedure, yielding a hierarchical, multi-layer road network.

The inputs to the procedure are a reference city $\mathcal{C}_0$ and a user specification $\mathcal{S}$. The reference city $\mathcal{C}_0$ serves as the calibration target and is represented as a compressed road-network graph. The user specification $\mathcal{S}$ is a natural-language prompt describing the intended structure of the target city. It may include qualitative descriptors (e.g., grid-like, radial, organic) as well as quantitative preferences, such as an acceptable calibration tolerance. The first stage generates a sequence of candidate cities $\mathcal{C}_k$ indexed by the iteration counter k. The loop starts at $k = 0$ and is bounded by the maximum number of iterations $k^{\max}$. Thus, k denotes the current iteration and $k^{\max}$ denotes the hard stopping criterion. The variable δ denotes the current calibration discrepancy between the reference and the candidate, i.e., $\delta = d(\mathcal{C}_0, \mathcal{C}_k)$, where $d(\mathcal{C}_0, \mathcal{C}_k) \geq 0$ is the city-to-city distance defined in Equation 1. In this article, d is operationalised as the one-dimensional Wasserstein distance between the empirical distributions of road-segment lengths in $\mathcal{C}_0$ and $\mathcal{C}_k$ (Equation 2), so δ measures how far the candidate's length distribution deviates from the reference distribution. At the beginning of the loop, we set $\delta = \emptyset$, which denotes the absence of feedback from a previous candidate. In subsequent iterations, δ becomes a numeric value that is used as structured feedback passed into the generator. The acceptance threshold is denoted by ε, and a candidate is considered sufficiently calibrated if $\delta \leq \varepsilon$.

At each iteration, the LLM agent invokes the generative routine `genHighLevelRoadNetwork`$(\mathcal{S}, \delta)$ to produce a candidate city $\mathcal{C}_k$. The term "LLM agent" refers to an LLM-driven controller that (a) synthesises a candidate high-level road structure consistent with $\mathcal{S}$ and (b) decides whether to regenerate based on tool-based validation results. Two tools are available to the agent. The tool `isPlanar`$(\mathcal{C}_k)$ returns a boolean value indicating whether the candidate road graph is planar, and the tool `wassersteinDistance` computes the calibration discrepancy $\delta = d(\mathcal{C}_0, \mathcal{C}_k)$. Planarity is treated as a hard feasibility requirement because the subsequent hierarchical refinement procedure relies on extracting planar regions

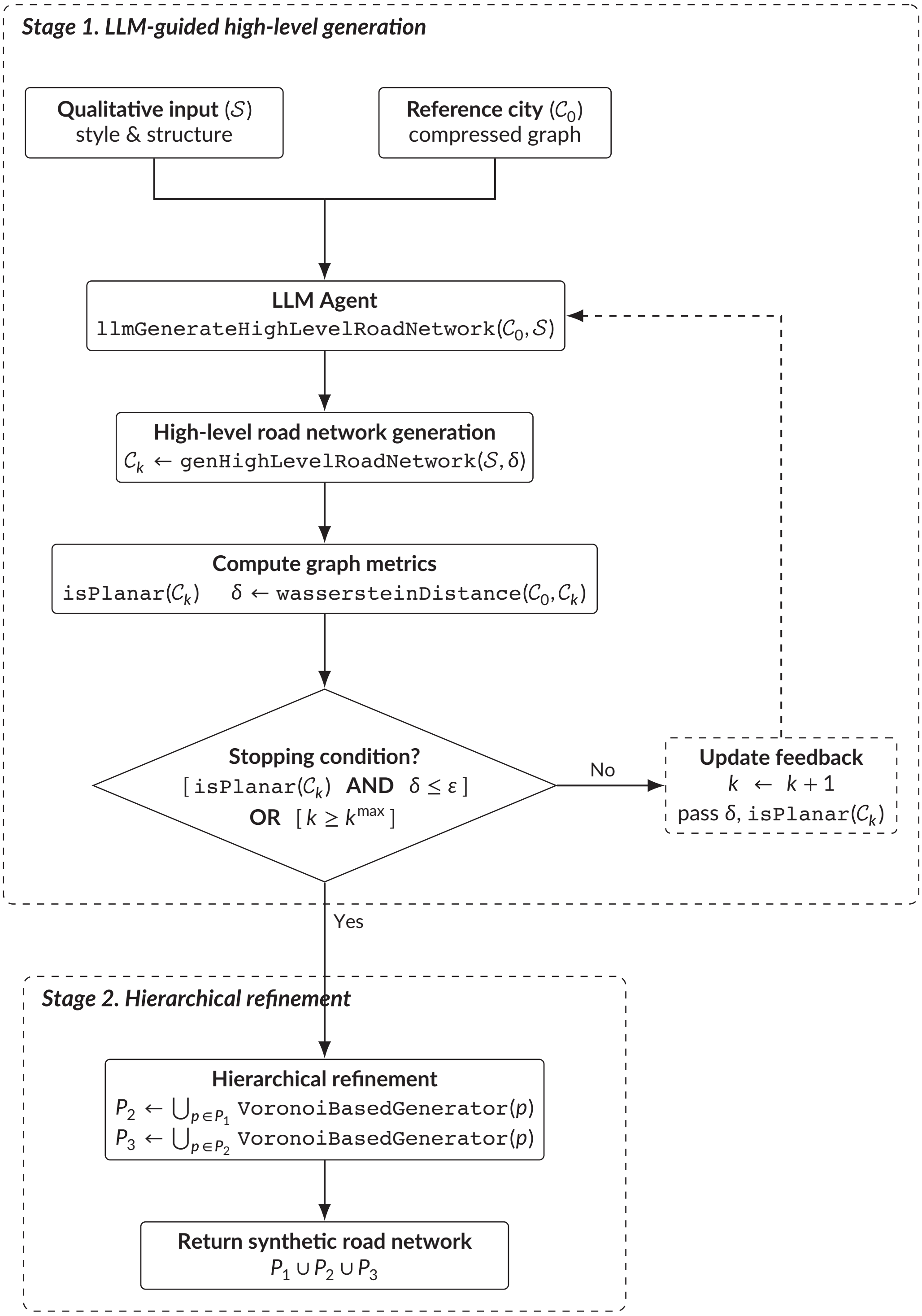

Figure 1. Flowchart of the proposed two-stage framework, where an LLM-guided iterative loop generates a planar high-level road network and a hierarchical Voronoi-based refinement adds lower-level streets.

(faces/polygons) from the road embedding; therefore, non-planar candidates are rejected by the stopping condition. The loop terminates when either the acceptance condition holds, i.e., when `isPlanar`($\mathcal{C}_k$) is true and $\delta \leq \varepsilon$, or when the iteration budget is exhausted, i.e., when $k \geq k^{max}$. The output of the first stage is the final accepted high-level candidate $\mathcal{C}_k$, which provides the arterial backbone for further refinement.

The second stage (shown in Figure 1) performs hierarchical refinement of the accepted high-level planar road network by populating lower-level streets within the planar regions induced by the backbone. Let P_1 denote the set of level-1 regions (faces) obtained from the planar subdivision defined by the high-level network returned by `llmGenerateHighLevelRoadNetwork`$(\mathcal{C}_0, \mathcal{S})$. For each region $p \in P_1$, the procedure `VoronoiBasedGenerator`(p) generates an internal lower-level street layout and returns a set of sub-regions, i.e., new polygons created by this refinement. Conceptually, the Voronoi-based generator samples seed points inside p, constructs the Voronoi tessellation induced by these seeds, and then converts selected cell boundaries into street segments, thereby partitioning p into smaller blocks. Aggregating these sub-regions across all level-1 regions yields the level-2 region set P_2, and applying an analogous refinement step once more to all regions in P_2 yields the level-3 region set P_3. The final multi-level city representation is the union $P_1 \cup P_2 \cup P_3$, which encodes the full three-level hierarchical subdivision and the corresponding street network generated within each level. The synthetic streets obtained at each level are assigned the corresponding road classes.

4. Numerical Experiments and Validation

We evaluate the LLM-backed synthetic city modelling method by conducting a set of experiments where the target was the main road network of Warsaw's Ochota district. The goal was to generate a synthetic high-level road graph that (a) matches a selected statistical property of the reference system—namely, the distribution of road-segment lengths—and (b) is planar, because planarity is required for extending the network with lower-level streets using a Voronoi-based procedure. In addition to this main experiment, we conducted supplementary analyses to assess the applicability of the framework to reference systems with different spatial organisations, its robustness across different LLM backends, and its sensitivity to the temperature parameter. We start by describing the experimental design (Section 4.1); next, we present the computational results (Section 4.2); and finally, we discuss the experimental outcomes (Section 4.3).

4.1. Experimental Design

The experiment tests whether the proposed LLM-centred framework can reproduce selected quantitative characteristics of a real-world district without copying its geometry. We use the Ochota district in central Warsaw (approximately 3 km in diameter) as the reference system. The intended outcome is therefore not a geometric replica of Ochota, but a structurally plausible alternative road network that matches the chosen calibration targets and can serve as a backbone for hierarchical refinement.

Road network data for Ochota were obtained from OpenStreetMap and converted into a weighted, undirected graph as described in Section 3.1. We retained only roads labelled with the OpenStreetMap tags "motorway," "trunk," "primary," and "secondary," treating them as the district's high-level transportation skeleton. The graph was then cleaned so that vertices correspond only to true junctions and endpoints: Intermediate shape points along road polylines were removed by compressing degree-2 paths into single weighted edges while preserving total segment lengths (Section 3.1.3). The resulting preprocessed graph constitutes the reference city $\mathcal{C}_0$ used for calibration.

To isolate the contribution of the LLM layer, we compare two generation settings that share the same reference city $\mathcal{C}_0$, the same acceptance thresholds, and the same evaluation pipeline: (a) a non-agentic baseline in which a high-level road network is generated using the Voronoi-based generator without iterative parameter updates (Bauer et al., 2010) and (b) an LLM-guided loop in which the full iterative procedure in Stage 1 is executed, i.e., the LLM produces candidate networks and uses tool-based feedback (`isPlanar`, `wassersteinDistance`) to refine its subsequent proposals until acceptance. For the LLM-guided experiments, we used Claude Opus 4.5 with temperature set to 1.0, top $p = 1.0$, a context window of 200,000 tokens, and a Wasserstein acceptance threshold of $\varepsilon = 0.05$. In the LLM-guided setting, we use a hard stopping criterion of $k^{\max} = 30$ iterations. If the acceptance conditions are not met within the budget, the best-so-far candidate is retained for reporting.

In both settings, we quantify similarity to the reference using the road-segment length distribution discrepancy defined in Section 3.1.2. Concretely, for each generated graph, we compute (a) the one-dimensional Wasserstein distance between the normalised empirical road-segment length distributions of the reference and the generated network (Equation 2) and (b) the two-sample Kolmogorov–Smirnov (KS) distance between the corresponding empirical cumulative distribution functions (ECDFs; Equation 3, defined in Section 4.2.1).

After obtaining an acceptable planar high-level graph, we apply the Voronoi-based subdivision procedure twice to generate two additional lower road levels, producing a three-level synthetic road network consistent with the hierarchical generation setup in Stage 2. These extended networks are used for qualitative illustration. To quantify variability, we run each setting for $n = 16$ independent trials. The next section reports the distributions of Wasserstein and KS distances, along with representative qualitative examples of generated networks.

4.2. Results

4.2.1. LLM-Based Approach and Voronoi-Based Approach Comparison

We evaluate whether the proposed LLM-centred loop (Stage 1) improves calibration of synthetic road networks relative to a non-agentic Voronoi-based baseline. In both settings, the reference city $\mathcal{C}_0$ is fixed, and each run produces a candidate high-level network $\mathcal{C}_k$. In the baseline setting, a single candidate $\mathcal{C}_1$ is generated directly by the Voronoi-based generator using fixed (default) settings (i.e., without iterative updates of generation parameters). In the LLM-guided setting, candidates are generated iteratively and updated using tool-based feedback, and the loop terminates when $\texttt{isPlanar}(\mathcal{C}_k)$ holds and $\delta \leq \varepsilon$, or when $k \geq k^{\max}$. For each generated graph, we compute the one-dimensional Wasserstein distance $d\left(\mathcal{C}_0, \mathcal{C}_k\right)$ (Equation 2) between the normalised empirical road-segment length distributions of the generated network and the reference Ochota graph. The implementation of the proposed framework, together with example prompts and selected response logs, is publicly available at https://github.com/AdamKasinski/synthetic_cities.

Using the Voronoi-based generator, we generated $n = 16$ independent graphs and computed the Wasserstein distance for each run. Figure 2 shows two sample synthetic cities generated after Stage 1 (blue) and after Stage 2 (red). The baseline achieved a mean Wasserstein distance of $\overline{W} = 0.105$ with a standard deviation of 0.0316.

With the LLM-guided generation, we executed the full LLM-guided loop (Stage 1) for $n = 16$ independent trials with a hard stopping criterion of $k^{\max} = 30$ iterations. In all 16 trials, the method returned a planar high-level graph within the iteration limit and produced the Wasserstein distances reported in Table 2. The number of iterations required for acceptance varied substantially across runs, ranging from 2 to 28, with a mean of 14. The observed variability was strongly affected by the conversational context in which the model was run.

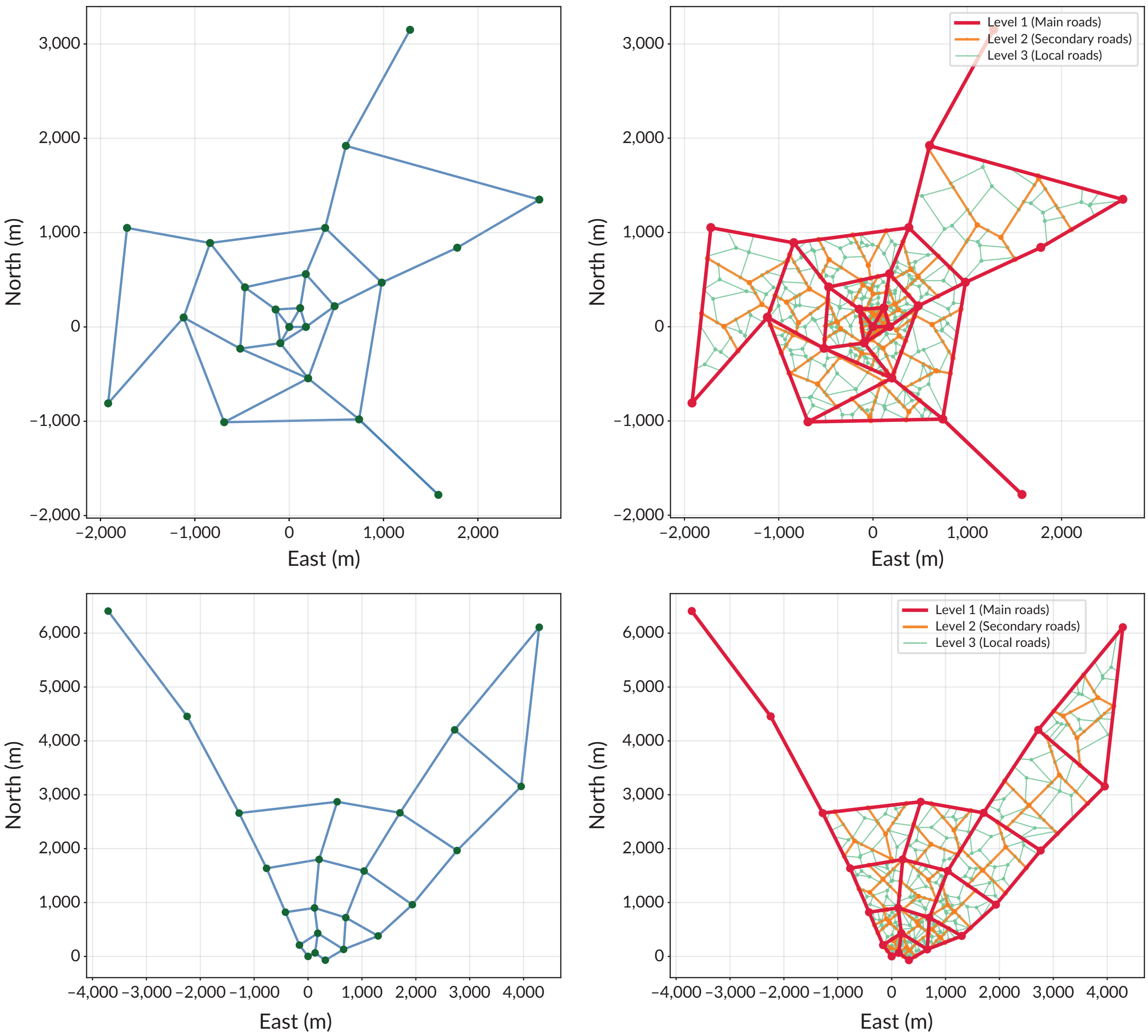

Figure 2. Two examples of the synthetic city generation process for Ochota District. The left column shows the planar high-level road backbones generated in Stage 1, and the right column shows the corresponding three-level networks obtained by two Voronoi-based refinement steps in Stage 2.

We also assessed distributional similarity using the two-sample KS distance, which captures the maximum pointwise deviation between ECDFs. Let $\hat{F}_0$ and $\hat{F}_k$ denote the ECDFs of the (normalised) road-segment length samples from the reference city $\mathcal{C}_0$ and a generated candidate $\mathcal{C}_k$, respectively. The KS distance is defined as:

$$D_{\mathrm{KS}}\left(\mathcal{C}_0, \mathcal{C}_k\right) = \sup_{x \in \mathbb{R}} \left|\hat{F}_0(x) - \hat{F}_k(x)\right| \tag{3}$$

Table 2. Wasserstein, KS, and Cramér–von Mises (CvM) distances for the 16 LLM-guided runs (planar and within $k^{max} = 30$ iterations).

Run	Wasserstein distance	KS distance	CvM distance
1	0.058	0.098	0.028
2	0.042	0.135	0.072
3	0.050	0.161	0.088
4	0.041	0.076	0.024
5	0.019	0.073	0.013
6	0.002	0.049	0.008
7	0.044	0.061	0.015
8	0.059	0.122	0.031
9	0.038	0.073	0.019
10	0.054	0.126	0.053
11	0.049	0.179	0.114
12	0.073	0.122	0.047
13	0.048	0.110	0.032
14	0.050	0.073	0.020
15	0.043	0.089	0.028
16	0.025	0.049	0.011
Mean	0.043	0.100	0.038
Std	0.016	0.039	0.030

Across $n = 16$ runs, the Voronoi baseline achieved a mean KS distance of $\overline{D}_{KS} = 0.156$ with standard deviation 0.026, whereas the LLM-guided procedure achieved $\overline{D}_{KS} = 0.100$ with standard deviation 0.039.

We also report the two-sample Cramér–von Mises (CvM) discrepancy, which captures the overall difference between the ECDFs. Across 16 runs, the Voronoi-based baseline achieved a mean CvM distance of 0.0881 (std. 0.0400), compared with 0.0377 (std. 0.0302) for the LLM-guided procedure, indicating a closer overall match to the road-segment length distribution. The following visual comparisons show two randomly selected examples from each method.

Figure 3 visualises the ECDFs of road-segment lengths for the reference Ochota network (solid blue) versus representative synthetic outputs (dashed red). The two upper panels (Voronoi-based) exhibit visibly larger vertical gaps between the ECDFs, indicating systematic distributional mismatch. Depending on the run, the Voronoi generator tends to over- or under-represent certain length ranges—most noticeably in the lower-to-mid segment lengths—which translates into a larger maximum ECDF deviation. In contrast, the two lower panels (LLM-guided) show the red curve closely tracking the blue curve across most of the support, with only small local departures.

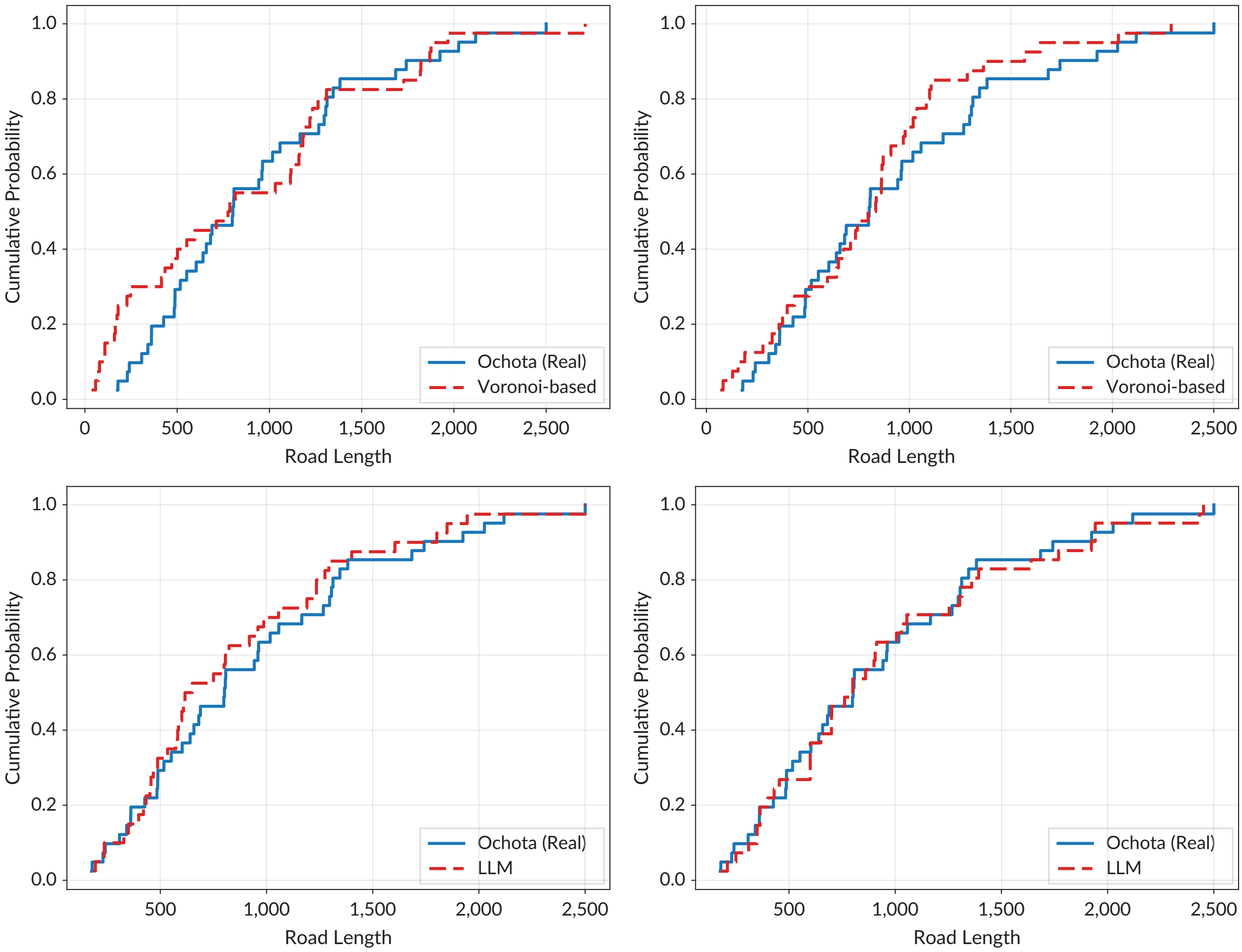

Figure 3. ECDFs of road-segment lengths for Ochota (blue) and synthetic networks (red). Top row: Voronoi-based; bottom row: LLM-guided.

Figure 4 shows the pointwise absolute differences between the ECDF of road-segment lengths in Ochota and the ECDF of each synthetic network, i.e., $\left|\hat{F}_{\text{Ochota}}(x) - \hat{F}_{\text{synthetic}}(x)\right|$. In the Voronoi-based examples, the difference curves exhibit higher and more frequent peaks, indicating larger local mismatches between the cumulative distributions over specific length ranges. In contrast, the LLM-guided examples display generally lower amplitudes and fewer pronounced peaks, meaning that the cumulative distribution of synthetic segment lengths tracks the reference more closely across most of the support.

While the absolute-difference curves quantify the magnitude of the mismatch, Figure 5 reports the corresponding signed ECDF differences, $\hat{F}_{\text{Ochota}}(x) - \hat{F}_{\text{synthetic}}(x)$, which reveal the direction of the deviation. Positive values indicate that Ochota accumulates probability mass faster up to length x—i.e., Ochota has relatively more short segments, so the synthetic network under-generates short segments up to that threshold. Negative values indicate the opposite—the synthetic ECDF is higher, meaning the synthetic network over-generates short segments up to x. In the representative Voronoi-based runs, the signed differences show larger and more persistent departures from zero across parts of the length range, suggesting systematic over-or under-generation in specific regimes, whereas the LLM-guided runs remain closer to zero, indicating fewer directional biases and a closer overall match to the reference distribution.

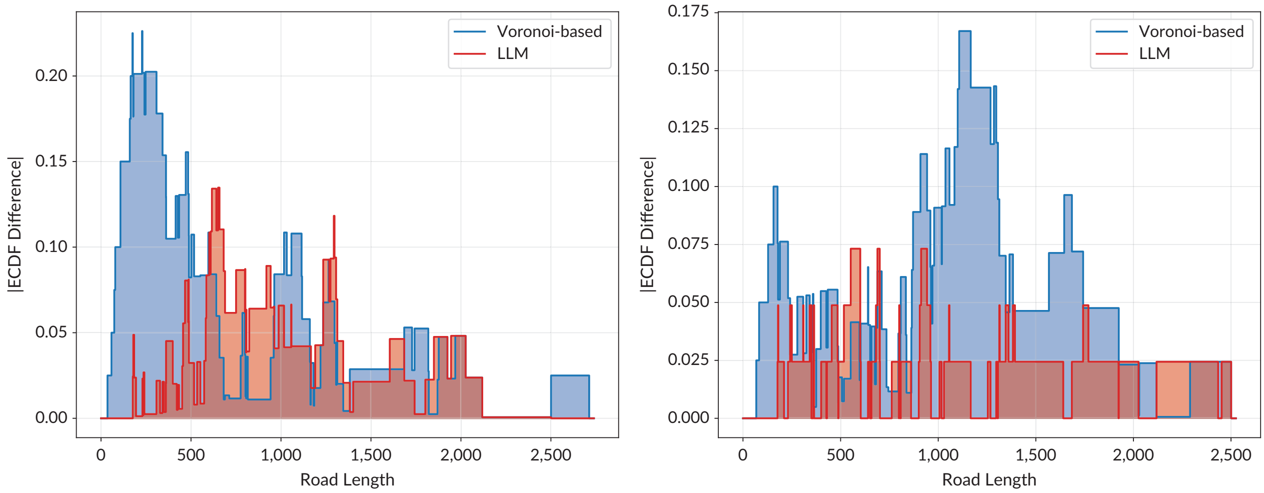

Figure 4. Overlay of absolute ECDF differences between Ochota and synthetic road-length distributions: Voronoi-based (blue) vs. LLM-guided (red), shown for two representative runs.

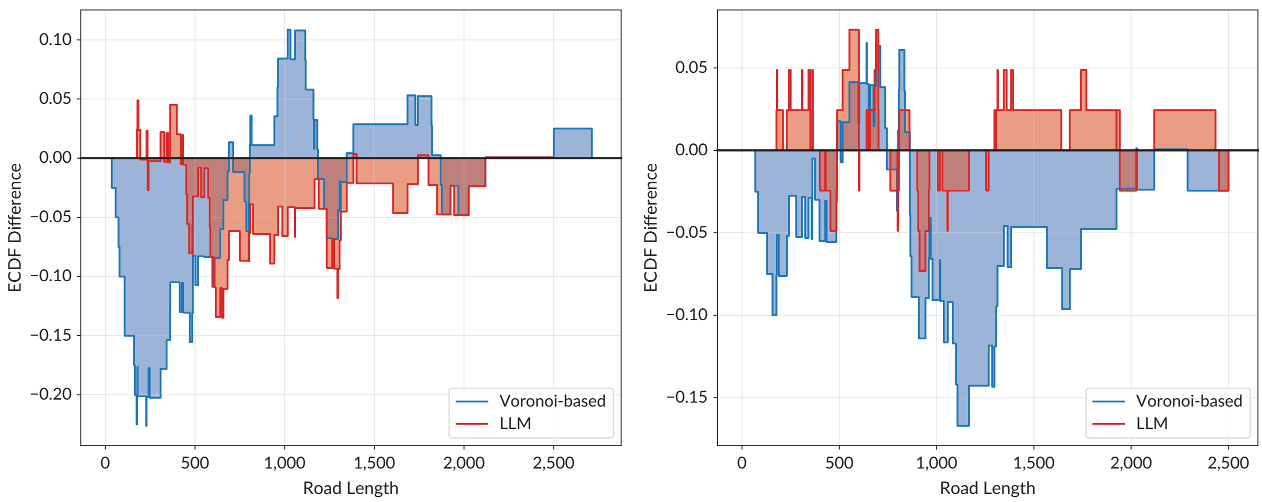

Figure 5. Overlay of ECDF differences between Ochota and synthetic road-length distributions: Voronoi-based (blue) vs. LLM-guided (red), shown for two representative runs.

Figures 6 and 7 summarise how the synthetic road-segment length distributions compare to the Ochota reference. In Figure 6, the Voronoi-based examples show several bins where the synthetic distribution visibly departs from Ochota, indicating local over- and under-representation of specific length ranges. The LLM-guided examples still differ across bins, but the most visually prominent mismatches are typically less pronounced in these representative runs. Figure 7 complements this by comparing quantiles directly: points closer to the $y = x$ line indicate better agreement across the full distribution, while systematic deviations from the diagonal reveal mismatches in particular length regimes (e.g., short or long segments). In the shown examples, the LLM-guided outputs track the reference quantiles more closely than the Voronoi-based baseline.

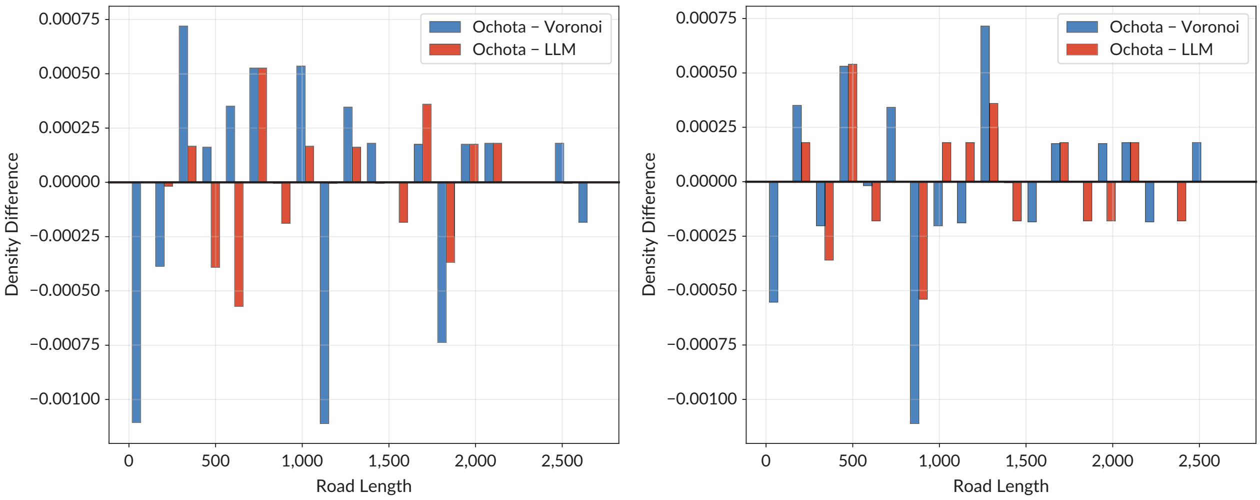

Figure 6. Differences between density histograms of road-segment lengths (Ochota minus synthetic), overlaid for Voronoi-based (blue) and LLM-guided (red), shown for two representative runs.

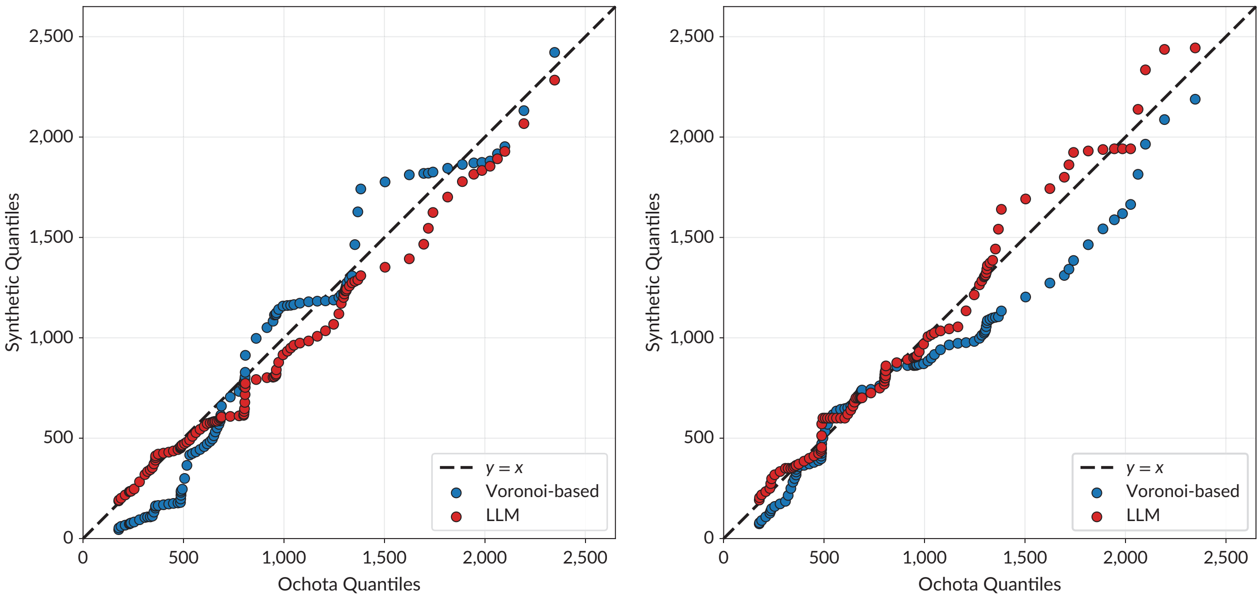

Figure 7. Q–Q plots of synthetic vs. Ochota road-segment lengths for two representative runs. The dashed line ($y = x$) indicates perfect agreement. Deviations from the diagonal indicate quantile-level mismatch.

This visual pattern is consistent with the quantitative KS results: The LLM-guided procedure reduces the maximum ECDF discrepancy (smaller KS distance), meaning it better matches not only the average segment length but also the overall distributional shape, including the relative mass of short, medium, and long road segments.

4.2.2. Results Across Different Reference Structures

To further assess applicability across urban structures, we generated synthetic cities for two additional reference areas: Vienna's Innere Stadt, representing a circular form, and Downtown Philadelphia, representing a grid-like structure. The synthetic examples were generated using the same procedure as in the Ochota experiment, so that the results remain directly comparable across cases. In both additional reference areas, all generated cities were planar and satisfied the same Wasserstein acceptance threshold of 0.05. The mean Wasserstein distances were 0.024 for Innere Stadt and 0.049 for Downtown Philadelphia, compared with 0.043 for Ochota. The corresponding KS values were 0.141, 0.153, and 0.100, while the CvM values were 0.115, 0.235, and 0.038, respectively. The weaker performance for Downtown Philadelphia should be interpreted in light of the greater size and structural complexity of the reference system, which contains approximately 70 road segments, compared to about 40 in the Ochota case. At the same time, the obtained discrepancies remained below the adopted Wasserstein threshold and were still better than those observed for random graphs of comparable size, which yielded mean values of 0.09 for Wasserstein, 0.40 for KS, and 1.73 for CvM. Figure 8 presents representative high-level road networks generated for the additional reference structures under different prompts at the first stage of the framework, before hierarchical refinement is applied, showing that the method can produce multiple variants while preserving the main structural characteristics of each reference layout. Taken together, these results indicate that the framework generalises beyond a single reference morphology, although calibration becomes more demanding for larger and more structurally complex road systems.

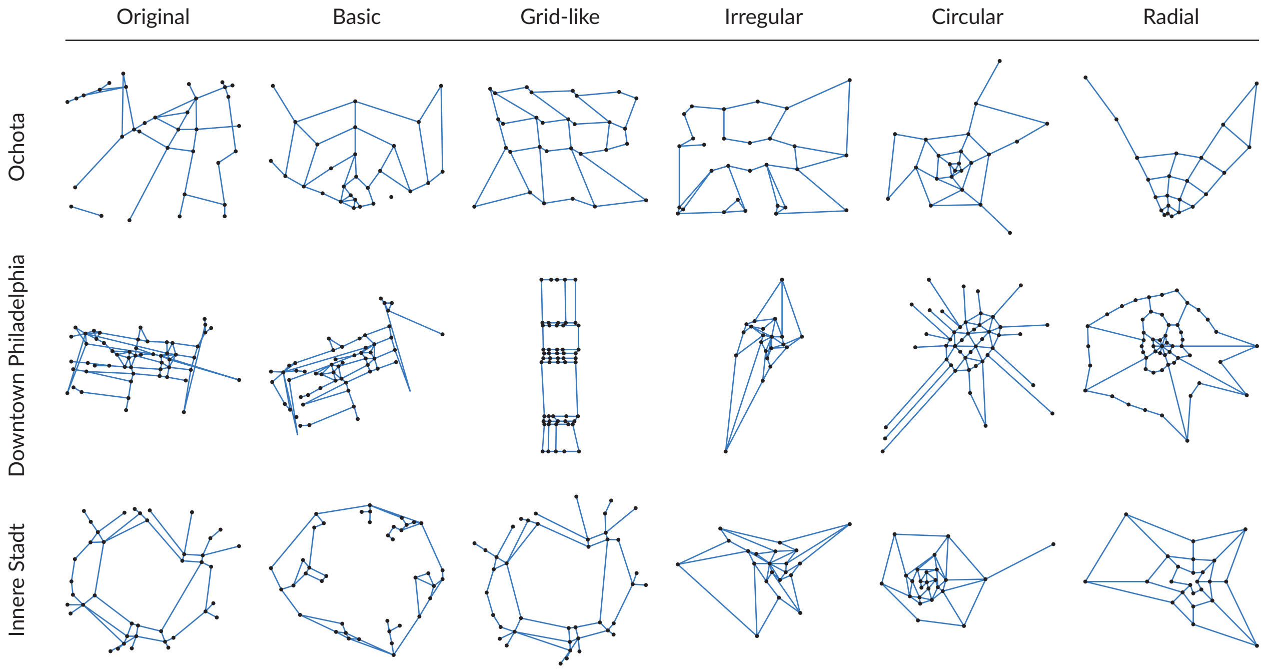

Figure 8. Synthetic road networks generated for three reference districts under five prompt types, illustrating the diversity of outputs obtained for each input layout.

4.2.3. Cross-Backend Comparison

To evaluate the robustness of the proposed method across different model families, we replicated the Ochota experiment using several LLM backends, namely Claude Opus 4.5, Gemini 3, Qwen 3.5, and Kimi K2.5. These supplementary experiments were intended to assess whether the proposed framework depends strongly on a specific model backend, or whether the same generation-and-validation procedure can be applied more broadly across different LLMs. All experiments were conducted under the same configuration, with the Ochota district as the reference structure, a Wasserstein acceptance threshold of 0.05, and identical remaining generation parameters. For each backend, we generated 16 synthetic cities. In all cases, the convergence rate to a valid output was 100%. The results show that the proposed framework remains effective across different LLM backends, although the relative ranking of models depends on the metric considered. Claude Opus 4.5 and Gemini 3 achieved the same mean Wasserstein distance of 0.043, but Claude produced lower KS and CvM discrepancies (0.100 and 0.038, respectively) than Gemini 3 (0.131 and 0.087). Qwen 3.5 achieved the lowest mean Wasserstein distance (0.016), yet this was accompanied by higher KS and CvM values (0.150 and 0.109), indicating a weaker match of the overall road-segment length distribution. By contrast, Kimi K2.5 achieved the lowest mean KS distance (0.091) while also maintaining competitive Wasserstein and CvM values (0.040 and 0.038). Taken together, these results indicate that the proposed framework is transferable across the tested LLM backends, in the sense that the same generation-and-validation procedure remains applicable across multiple models. At the same time, backend choice affects which aspect of distributional similarity is matched most closely.

4.2.4. Temperature Sensitivity Analysis

In the main experiments, the temperature parameter was set to 1.0. This choice was motivated by the exploratory character of the proposed framework. Because the method operates within an iterative generation-and-validation loop, a higher temperature was intended to encourage the model to explore a broader range of candidate network structures rather than produce overly deterministic outputs, while the external validation procedure—including the planarity check and the Wasserstein-distance-based feedback—acts as a stabilising mechanism that filters out invalid or poorly calibrated candidates. To assess the sensitivity of the framework to this parameter, we conducted a controlled temperature analysis. Due to technical constraints of the command line interface environment used for the primary Claude experiments, the temperature parameter could not be directly modified in that setting. We therefore repeated the Ochota experiment using the Gemini backend and evaluated the framework at four temperature values: $T = 1.0$, 0.5, 0.2, and 0.0. The results suggest that the proposed framework is only moderately sensitive to temperature within the tested stochastic range. The mean Wasserstein distances for $T = 1.0$, 0.5, and 0.2 were relatively similar, at 0.043, 0.046, and 0.045, respectively. The most pronounced deterioration was observed for the deterministic setting $T = 0.0$, for which the mean Wasserstein distance increased to 0.089. The KS and CvM values also varied across temperature settings, but they did not reveal a strong or monotonic relationship. Overall, these results indicate that temperature influences the outputs of the framework, but that the differences among stochastic settings remain limited, whereas fully deterministic generation performs less favourably.

4.3. Discussion

The results show that the LLM-guided procedure produces high-level road networks whose road-segment length distributions are closer to the reference layout than those generated by the non-agentic Voronoi baseline. Using Wasserstein distance as the primary calibration measure, the LLM-guided runs achieved a lower mean discrepancy ($\overline{W} = 0.043$, std. 0.016, $n = 16$) than the baseline ($\overline{W} = 0.105$, std. 0.0316, $n = 16$), corresponding to an approximately 2.5-fold reduction in mean error. The same pattern appears for the KS distance, whose mean decreased from $\overline{D}_{KS} = 0.156$ (std. 0.026) for the baseline to $\overline{D}_{KS} = 0.100$ (std. 0.039) for the LLM-guided runs. Taken together, these results indicate that the iterative LLM-guided loop more consistently steers generation towards candidates whose segment-length statistics align with the reference distribution.

In all 16 runs, the LLM-guided procedure returned a planar high-level graph within the iteration limit. This is important because the subsequent hierarchical extension relies on extracting faces and refining them with the Voronoi-based subdivision. The qualitative examples in Figure 2 show that the accepted planar backbones can be extended into coherent multi-level synthetic networks.

Additional experiments broaden these findings in three directions. First, applying the framework to Vienna's Innere Stadt and Downtown Philadelphia showed that it can generate planar synthetic cities for reference systems with markedly different spatial organisation, although calibration becomes more demanding for larger and more structured networks. Second, experiments with Gemini 3, Qwen 3.5, and Kimi K2.5 showed that the same generation-and-validation procedure remains applicable across multiple LLM backends, even though the relative ranking of models depends on the metric considered. Third, the temperature-sensitivity analysis indicated only moderate sensitivity within the tested stochastic range, while the deterministic setting performed less favourably.

At the same time, the current framework uses a simplified high-level representation to keep the structural description concise and easier for the LLM to process reliably. This representation can subsequently be extended in a post-processing step to obtain a more realistic transport network. In particular, each undirected edge may be converted into one or two directed edges to represent one-way and two-way streets, while the hierarchical information already used in the model can serve as a basis for assigning road classes and related attributes. Additional restrictions, such as turn limitations at intersections, may then be introduced probabilistically, with their likelihood depending on the classes of the intersecting roads. The relevant parameters could be estimated from spatial or transport datasets, allowing the generated structures to be enriched with more realistic operational detail without changing the core generation procedure.

5. Conclusion

This article presents a new approach to synthetic city generation that uses LLMs to connect high-level natural-language requirements with quantitative validation. The proposed framework has two stages. First, an LLM-guided iterative loop generates a planar, high-level road network (the arterial backbone) and refines successive candidates using tool-based feedback. Second, the accepted planar backbone is hierarchically refined using a Voronoi-based procedure that generates lower-level streets within the planar regions induced by the backbone, yielding a multi-level road network.

We demonstrate the method's applicability on a real-world OpenStreetMap dataset. Numerical experiments show that the proposed method improves similarity to the reference city relative to a non-agentic Voronoi-based baseline. When calibration is defined by the road-segment length distribution, the LLM-guided runs achieved lower discrepancies than the baseline in both Wasserstein distance (mean 0.043 vs. 0.105 over $n = 16$ runs) and KS distance (mean 0.100 vs. 0.156). In addition, all LLM-guided runs produced planar high-level graphs, enabling the downstream hierarchical extension step based on planar subdivision.

Supplementary experiments further showed that the framework remains applicable across reference systems with different spatial organisation and across multiple LLM backends, although calibration becomes more demanding for larger and more structured networks and model rankings depend on the evaluation metric.

Temperature sensitivity was moderate within the tested stochastic range, while fully deterministic generation performed less favourably.

The main advantage of the proposed method is that it simultaneously exhibits four desired properties that are largely missing from existing approaches: It supports calibration to a reference city, enforces conditional constraints such as planarity and a similarity tolerance, supports zero-shot customisation through new prompts and reference inputs, and uses semantic reasoning to translate qualitative requirements into measurable checks. In this way, we demonstrate that LLMs can be incorporated into the synthetic city generation process and that doing so improves the quality of the generated cities under the adopted evaluation metrics.

The current evaluation focuses on a single calibration target: the normalised distribution of road-segment lengths. Future work should extend the calibration objective to include additional complementary transport-network characteristics (e.g., intersection density and degree statistics, orientation distributions, circuity, or routing-based measures such as APLS), and investigate multi-objective calibration in which several metrics are optimised simultaneously. Another important direction for future work is to enrich the simplified high-level graph representation with additional transport attributes, such as directionality, road classes, capacities, or turn restrictions, thereby improving the realism of the generated networks and broadening their applicability in simulation studies.

Funding

This research was funded by the National Science Centre (NCN), Poland (grant number 2021/41/B/HS4/03349).

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The source code and supplementary materials supporting this study, including the implementation of the proposed framework, example prompts, and selected response logs, are openly available in the public repository at https://github.com/AdamKasinski/synthetic_cities.

LLMs Disclosure

During the preparation of this work, the authors used the LLM models Overleaf Writefull and OpenAI's ChatGPT to improve the readability and language of the manuscript, correct grammatical and spelling errors, and enhance overall language quality and clarity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Abraham, I., Fiat, A., Goldberg, A. V., & Werneck, R. F. (2010). Highway dimension, shortest paths, and provably efficient algorithms. In M. Charikar (Ed.), *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 782–793). Society for Industrial and Applied Mathematics.

Almirall, E., Callegaro, D., Bruins, P., Santamaría, M., Martínez, P., & Cortés, U. (2022). *The use of synthetic data to solve the scalability and data availability problems in smart city digital twins*. arXiv. https://doi.org/10.48550/arXiv.2207.02953

Barbie, A., & Hasselbring, W. (2024). *Toward reproducibility of digital twin research: Exemplified with the PiCar-X*. arXiv. https://doi.org/10.48550/arXiv.2408.13866

Barthelemy, M. (2019). The statistical physics of cities. *Nature Reviews Physics*, *1*, 406–415.

Batty, M., & Hudson-Smith, A. (2001). Virtuality and cities: Definitions, geographies, designs. In P. F. Fisher & D. Unwin (Eds.), *Virtual reality in geography* (pp. 270–291). Taylor & Francis.

Bauer, R., Krug, M., Meinert, S., & Wagner, D. (2010). Synthetic road networks. In B. Chen (Ed.), *Algorithmic Aspects in Information and Management: 6th International Conference, AAIM 2010* (pp. 46–57). Springer.

Beneš, J., Wilkie, A., & Křivánek, J. (2014). Procedural modelling of urban road networks. *Computer Graphics Forum*, *33*(6), 132–142.

Bilal, S., Zaatour, W., Otano, Y. A., Saha, A., Newcomb, K., Kim, S., Kim, J., Ginjala, R., Groen, D., & Michael, E. (2025). CitySEIRCast: An agent-based city digital twin for pandemic analysis and simulation. *Complex & Intelligent Systems*, *11*(1), Article 83.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(Suppl. 3), 7280–7287.

Borgwardt, K. M., & Kriegel, H.-P. (2005). Shortest-path kernels on graphs. In J. Han, B. W. Wah, V. Raghavan, X. Wu, & R. Rastogi (Eds.), *Fifth IEEE International Conference on Data Mining (ICDM'05)* (pp. 74–81). IEEE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chen, G., Esch, G., Wonka, P., Müller, P., & Zhang, E. (2008). Interactive procedural street modeling. In *ACM SIGGRAPH 2008 Papers* (Article 103). Association for Computing Machinery.

Chen, J., Xu, W., Cao, H., Xu, Z., Zhang, Y., Zhang, S., Zhang, Z., & Yu, B. (2026). Network generation artificial intelligence: Multimodal road network generation based on large language models. *Transportation Research Record*, *2680*(1), 432–451.

Chu, H., Li, D., Acuna, D., Kar, A., Shugrina, M., Wei, X., Liu, M.-Y., Torralba, A., & Fidler, S. (2019). Neural turtle graphics for modeling city road layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4522–4530). IEEE.

Corcoran, P., & Jones, C. B. (2022). A persistent homology model of street network connectivity. *Transactions in GIS*, *26*(1), 155–181.

Feng, M., & Porter, M. A. (2020). Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence. *Physical Review Research*, *2*(3), Article 033426.

Fortune, S. (1986). A sweepline algorithm for Voronoi diagrams. In *Proceedings of the Second Annual Symposium on Computational Geometry* (pp. 313–322). Association for Computing Machinery.

Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., Del Olmo, R., López-Paredes, A., & Edmonds, B. (2009). Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, *12*(1), Article 1.

Gan, W., Zhao, Z., Wang, Y., Zou, Y., Zhou, S., & Wu, Z. (2024). UDGAN: A new urban design inspiration approach driven by using generative adversarial networks. *Journal of Computational Design and Engineering*, *11*(1), 305–324.

Geisberger, R., Sanders, P., Schultes, D., & Delling, D. (2008). Contraction hierarchies: Faster and simpler hierarchical routing in road networks. In C. C. McGeoch (Ed.), *Experimental Algorithms: 7th International Workshop, WEA 2008* (pp. 319–333). Springer.

Hartmann, S., Weinmann, M., Wessel, R., & Klein, R. (2017). StreetGAN: Towards road network synthesis with generative adversarial networks. In V. Skala (Ed.), *Short Papers Proceedings of WSCG 2017* (pp. 133–142). Václav Skala – Union Agency.

Kim, J.-S., Kavak, H., & Crooks, A. (2018). Procedural city generation beyond game development. *SIGSPATIAL Special*, *10*(2), 34–41.

Li, W., Gu, J., Liu, S., Zhu, Y., Deng, S., Zhao, L., Han, J., & Cai, X. (2014). *Optimal transport in worldwide metro networks*. arXiv. https://doi.org/10.48550/arXiv.1403.7844

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). *On LLMs-driven synthetic data generation, curation, and evaluation: A survey*. arXiv. https://doi.org/10.48550/arXiv.2406.15126

Lüscher, P., & Weibel, R. (2013). Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems*, *37*, 18–34.

Neira, M., & Murcio, R. (2022). *Graph representation learning for street networks*. arXiv. https://doi.org/10.48550/arXiv.2211.04984

Nishida, G., Garcia-Dorado, I., & Aliaga, D. G. (2016). Example-driven procedural urban roads. *Computer Graphics Forum*, *35*(6), 5–17.

Papyshev, G., & Yarime, M. (2021). Exploring city digital twins as policy tools: A task-based approach to generating synthetic data on urban mobility. *Data & Policy*, *3*, Article e16.

Parish, Y. I. H., & Müller, P. (2001). Procedural modeling of cities. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 301–308). Association for Computing Machinery.

Rasal, S., & Boddhu, S. K. (2024). Beyond segmentation: Road network generation with multi-modal LLMs. In K. Arai (Ed.), *Intelligent Computing: Proceedings of the 2024 Computing Conference, Volume 1* (pp. 308–315). Springer.

Roth, J., Martin, A., Miller, C., & Jain, R. K. (2020). SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods. *Applied Energy*, *280*, Article 115981.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., & Borgwardt, K. (2009). Efficient graphlet kernels for large graph comparison. In D. van Dyk & M. Welling (Eds.), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (pp. 488–495). Proceedings of Machine Learning Research.

Song, A., & Whitehead, J. (2019). TownSim: Agent-based city evolution for naturalistic road network generation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games* (Article 81). Association for Computing Machinery.

Sun, Y., & Dogan, T. (2023). Generative methods for urban design and rapid solution space exploration. *Environment and Planning B: Urban Analytics and City Science*, *50*(6), 1577–1590.

Therias, A., & Rafiee, A. (2023). City digital twins for urban resilience. *International Journal of Digital Earth*, *16*(2), 4164–4190.

Togninalli, M., Ghisu, E., Llinares-López, F., Rieck, B., & Borgwardt, K. (2019). Wasserstein Weisfeiler–Lehman graph kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 6436–6446). Curran Associates.

van Vliet, J. (2013). *Calibration and validation of land-use models* [Unpublished doctoral dissertation]. Wageningen University.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Weber, B., Müller, P., Wonka, P., & Gross, M. (2009). Interactive geometric simulation of 4D cities. *Computer Graphics Forum*, *28*(2), 481–492.

Xu, H., Omitaomu, F., Sabri, S., Zlatanova, S., Li, X., & Song, Y. (2024). Leveraging generative AI for urban digital twins: A scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Informatics*, *3*(1), Article 29.

Yu, Q., & Steed, A. (2012). Example-based road network synthesis. In C. Andújar & E. Puppo (Eds.), *Eurographics 2012—Short Papers* (pp. 53–56). The Eurographics Association.

Zhang, X., & Shi, X. (2022). A method for road network generation based on tensor field and multi-agent. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *48*, 205–210.

Zheng, Y., Lin, Y., Zhao, L., Wu, T., Jin, D., & Li, Y. (2023). Spatial planning of urban communities via deep reinforcement learning. *Nature Computational Science*, *3*(9), 748–762.

About the Authors

Adam Kasiński is a PhD student at SGH Warsaw School of Economics. His research spans urban modelling and econometric modelling, with a particular focus on market risk analysis. His current work investigates how road network morphology shapes access to multimodal urban transport systems.

Przemysław Szufel is an associate professor at SGH Warsaw School of Economics and an adjunct professor at Toronto Metropolitan University. His research spans transportation modelling, numerical optimisation, and agent-based simulation. A devoted Julia language enthusiast, he holds 2nd place worldwide on StackOverflow for Julia answers and uses it daily in his computational work.