Agentic Urban Planning Assistant: A Parametric Approach to Spatial and Regulatory Evaluation in Urban Planning

Emil Hristov, Angel Spasov, Denis Dimitrov, Dessislava Petrova-Antonova

GATE Institute – Big Data for Smart Society, Sofia University "St. Kliment Ohridski", 5 James Bourchier Blvd., 1164 Sofia, Bulgaria
(emil.hristov@gate-ai.eu, an.spasov@gmail.com, denis.dimitrov@gate-ai.eu, dessislava.petrova@gate-ai.eu)

Keywords: Urban planning assistant, regulatory retrieval, cadastral data, spatial decision support, LangGraph

Abstract

Urban planning decision support often requires a joint interpretation of legal rules, parcel geometry, zoning parameters, street context, and proposal-specific constraints. Generic retrieval-augmented generation is insufficient for this setting since many answers depend on structured spatial evidence and deterministic regulatory calculations rather than on retrieved prose alone. This paper presents an agentic urban planning assistant designed for parcel-level first-pass analysis. It combines deterministic request classification, route-specific retrieval, parcel lookup, exact-index zoning-parameter resolution, parcel-side normalization, setback-rule assignment, gated planning calculations, and final large-language-model (LLM) answer synthesis. The implementation uses the LLM mainly for the final answer generation stage, while the evidence and reasoning foundation is constructed through structured data preparation and deterministic workflow logic. An evaluation framework is developed that aligns benchmark design and metric selection with the actual assistant's route structure. It combines deterministic checks, complementary LLM-as-a-judge metrics, benchmark generation from reviewed seed cases, and execution-level observability. A reviewed 40-case architecture-aware benchmark for the Sofia/Bulgarian case study is prepared. On this benchmark, the deterministic pass rate reached 1.00, while the complementary judge-based scoring yielded a faithfulness value of 0.931 under a revised structured-evidence evaluation setup. The remaining weaknesses are concentrated in proposal-feasibility synthesis and parcel-zoning descriptive completeness.

1. Introduction

Data-driven urban planning requires a joint evaluation of legal constraints, spatial characteristics, and planning objectives. Urban planners, investors, and public authorities must assess whether a proposed development is compliant with regulations, feasible within spatial constraints, and aligned with the applicable planning framework. The recent transition towards automated and agentic planning has laid the groundwork for parametric spatial planning. Adversarial frameworks have been applied to reimagine city configurations, enabling generative models to produce city layouts and incorporate real-world complexity (Wang et al., 2020). Automated urban planning leveraging deep reinforcement learning has been used to generate a functional city by optimising land-use configurations and layouts subject to spatial and infrastructure constraints (Park et al., 2023). Subsequent work introduces multi-objective optimizations balancing competing criteria (Zhou et al., 2024). UrbanGPT integrates a spatio-temporal dependency encoder with an instruction-tuning paradigm.

Parallel research has also advanced the integration of regulatory zoning knowledge into spatial AI systems, which provides the groundwork for rule-aware generative planning (Liu et al., 2025a). Smart city frameworks have additionally combined sensor-driven and parametric methods for real-time urban analysis (Kalyuzhnaya et al., 2025). Even more recently, LLM-agents have been integrated as planning assistants capable of interpreting spatial queries and proposing interventions of the planning (Wang et al., 2026). Retrieval-Augmented Generation (RAG) can also serve as a "trust anchor" by retrieving hard, source-linked evidence from a knowledge graph within a digital city twin. Such a framework is proposed to address complex urban inquiries grounded in verifiable facts (Xie et al., 2026). RAG has also been integrated into Intelligent Transportation Systems (ITS). First, an overview of the state of the art in ITS and CV is provided to support the rationale for an RAG. A conceptual multi-agent framework is proposed for smart mobility services in urban communities. In the concept, many challenges are outlined, including optimised task coordination, data sovereignty, and AI accountability (Xu et al., 2024).

A domain-specific question-answering GeoRAG was proposed, featuring a knowledge-enhanced QA method aligned with a geographic domain perspective. First, a corpus of material was collected, and then a Bert-Based multi-label classifier was added to specify different query types and construct retrieval evaluation pairs. GeoPrompt templates were designed to improve response quality, and competitive experiments were conducted for GeoRAG and traditional RAG implementations (Chen et al., 2026). Another use of a RAG in the Smart city domain, for transportation scheduling, environmental data analysis, and professional consultation services, was described in (Ku et al., 2025). A conceptual framework is proposed for an RAG that employs strategic engineering principles to enhance planning simulations. The framework uses agents within the simulation software to model various development scenarios (Bruzzone et al., 2024). For public transport, a TransRAG framework was proposed that integrates embedding-based retrieval with tree and graph structures to enhance both document indexing and question-answering performance, and the results show that it outperforms other RAG implementations (Liu et al., 2025b). A tourism recommendation RAG was compared to a generic LLM, as an agentic planner, to eliminate hallucination. The RAG was evaluated by researchers using 50 simulated travel personas. It achieved better Faithfulness and Relevancy and completely eliminated unsafe recommendations (Na Takuatung and Bussracumpakorn, 2026). For governance evaluation using RAG, reviewing the low-carbon practices of 296 Chinese cities (Cao et al., 2026).

Building on previous work, we argue that there is a gap in parcel-level planning assistance in urban planning. RAG technologies integrates specific domain knowledge about law regulation intersecting with the urban fabric. This paper is narrower the problem, proposing more operational than general urban AI assistance. The potential of LLMs for spatio-temporal learning is highlighted, particularly in zero-shot scenarios where labelled data is scarce (Li et al., 2024). The normative dimensions of algorithmic planning have also been examined, integrating equity and justice into urban workflows (Salazar-Miranda and Talen, 2025). The objective is first-pass parcel-level regulatory evaluation, where correctness, retrieval precision, route transparency, and calculation traceability are essential. This motivates a workflow in which routing is explicit, spatial interpretation and regulatory calculations are deterministic where possible, and answer synthesis follows from assembled evidence rather than replacing it. Unlike systems centred on scenario generation or general regulatory question answering, our approach focuses on parcel-level validation through route-aware execution and architecture-aware evaluation.

2. Geospatial Data and Regulatory Sources

The proposed assistant was developed and evaluated using a pilot area in Sofia, Bulgaria, as a real-world case study. Figure 1 illustrates the spatial datasets used in the study at three representative scales, from the pilot area to an individual cadastral parcel. It integrates heterogeneous spatial and regulatory datasets to support parcel-level planning evaluation. Structured spatial data includes cadastral parcels, buildings, streets, and zoning information, accessed primarily through relational and spatial queries. Regulatory sources consist of legal documents and a structured zoning annex containing the planning parameters required for parcel-specific evaluation. Depending on the question, the workflow retrieves legal provisions, spatial context, or both before performing deterministic planning calculations.

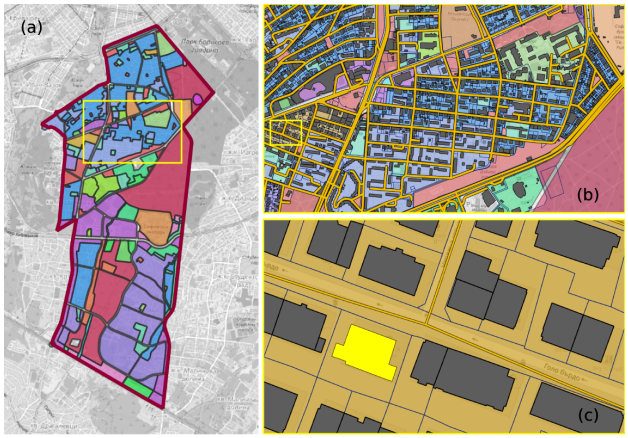

Figure 1. Multi-scale spatial datasets. (a) Pilot study area with Master Plan zoning. (b) Representative neighbourhood showing integrated zoning, street, building, and cadastral datasets. (c) Parcel-level view highlighting the selected building. Yellow rectangles indicate progressive zoom between spatial scales.

The combination of these sources, summarized in Table 1, is fundamental to the proposed workflow. Legal text alone cannot determine parcel-specific planning quantities, while spatial data alone cannot provide the regulatory basis required for compliant decision support. The assistant, therefore, integrates both through route-dependent retrieval and deterministic reasoning.

Source	Representation	Main role	Examples
Parcels	Relational geometry	Parcel identity and area context	KI/UPI lookup, area, proposal context
Buildings	Geometry and attributes	Existing built context	Footprint, GFA, building presence
Streets	Spatial tables with functional attributes	Frontage and setback interpretation	Street-facing side, street-class setbacks
Spatial zoning	Parcel-linked zoning fields	Zoning index and parcel context	Exact-index retrieval, profile selection
Legal text	Article-level indexed chunks	General regulatory retrieval	Legal explanation, reference retrieval
Zoning annex	Structured table rows	Parcel-specific parameter resolution	FAR, density, green area, cornice height

Table 1. Main data and regulatory sources.

Preparing the regulatory corpus is a core component of the proposed workflow rather than a simple preprocessing step. Regulatory documents are distributed as presentation-oriented HTML rather than retrieval-ready data, requiring the extraction of legal content, preservation of document hierarchy, and transformation into structured representations suitable for retrieval and parameter resolution.

Legal documents are converted into article-level representations while preserving the original hierarchy of chapters, sections, and articles. Each article is stored together with its structural metadata and textual content, which are combined to form the embedding representation. Article-level chunking preserves legal semantics and enables citation-aligned retrieval. The zoning annex is processed separately because it represents a structured hierarchical table rather than a narrative text. During preparation, the table is reconstructed, hierarchical relationships are preserved, and operational planning parameters such as zone labels, maximum density, FAR, green area, and cornice height are extracted together with their parent context.

The resulting corpus supports hybrid indexing. Narrative legal documents are embedded at the article level, while zoning regulations are embedded at the row level with preserved hierarchical context. During retrieval, this structure enables both precise parameter resolution and legally grounded explanations.

3. Agentic System Architecture

The proposed assistant is implemented as an explicit LangGraph workflow rather than as a single prompting loop. Request normalization, route classification, parcel lookup, retrieval, calculation gating, deterministic calculations, and final answer synthesis are implemented as separate workflow stages connected by conditional branches. This modular design improves transparency, inspection, and evaluation compared with monolithic retrieval-and-generation pipelines.

Figure 2 provides a simplified overview of the proposed route-aware workflow. Requests are first normalized and classified

into one of three execution routes. Parcel-specific requests trigger structured spatial retrieval and deterministic processing before state-aware answer synthesis, whereas general regulatory and legal-reference questions follow lighter retrieval paths.

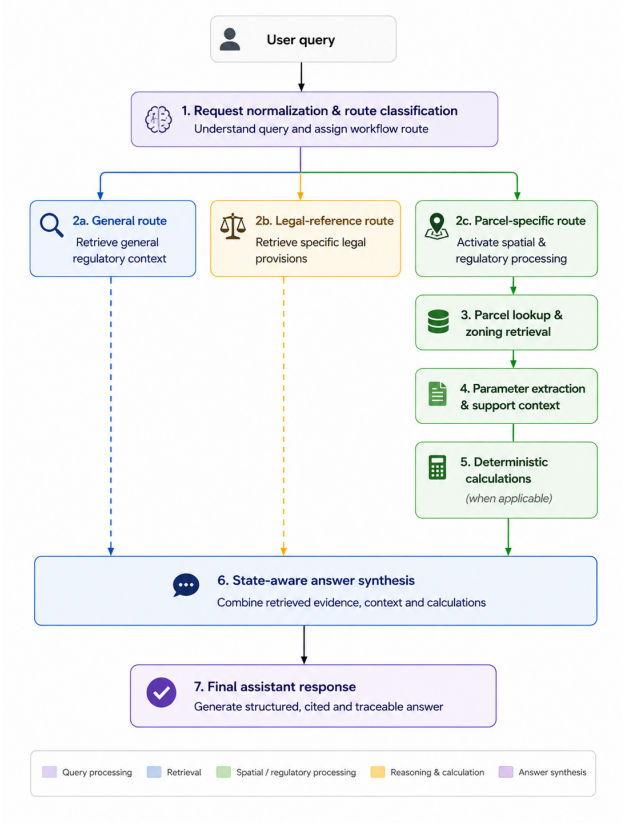

Figure 2. Simplified route-aware workflow.

A key design decision is that the initial routing stage is deterministic. Parcel identifiers, proposal intent, legal references, and calculation requests are extracted using explicit rules before any language model is invoked. The workflow supports both cadastral identifiers (KI) and anchored UPI–quarter references, including forms such as ‘‘UPI I, kv. 186’’ and ‘‘UPI XII-45’’, together with their Bulgarian-script anchored equivalents. The extractor deliberately rejects bare unanchored forms such as ‘‘I-186’’, which may refer to multiple entities and therefore cannot be resolved reliably. The LLM is invoked only after route-specific evidence has been assembled.

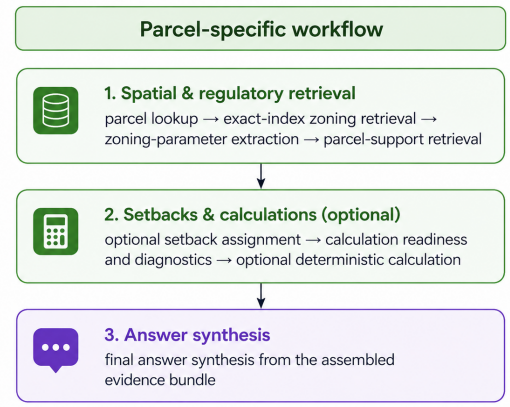

Figure 3. Simplified parcel-specific workflow.

Before database access, the workflow distinguishes resolved, missing, incomplete, invalid, and not-found parcel references. Input-quality issues are handled deterministically and returns predefined responses without triggering retrieval or final answer generation, whereas not-found cases are identified only after a valid parcel lookup.

The parcel-specific execution path is illustrated in Figure 3. The workflow distinguishes four main route families:

1. out-of-domain questions, which bypass legal retrieval and calculation;
2. general in-domain questions, which use broad legal retrieval;
3. general in-domain questions with explicit legal locators, which use a dedicated legal-reference branch;
4. parcel-specific questions, which combine parcel lookup, exact-index zoning retrieval, supporting legal retrieval, optional setback assignment, and conditional deterministic calculations.

This modular design makes routing, retrieval, deterministic reasoning, and answer synthesis explicit, supporting transparent execution, inspection, and architecture-aware evaluation.

4. Retrieval and Deterministic Reasoning

4.1 Route-Specific Retrieval Families

The retrieval strategy is route-specific rather than uniform. General in-domain questions without explicit legal references use broad semantic retrieval over the indexed legal corpus. Questions containing explicit legal locators follow a dedicated legal-reference branch that first narrows the candidate set using structured legal markers before applying semantic ranking and reranking. Consequently, evaluation of legal-reference retrieval should be performed on the final reranked evidence set rather than on the broader initial candidate pool.

Parcel-specific retrieval follows a different strategy. Instead of searching the full regulatory corpus, the workflow first uses the parcel zoning index to perform an exact lookup in the zoning annex associated with Article 3(2). If a single active row matches the zoning index, it is selected directly. If multiple rows share the same index, semantic ranking is applied only within that subset. This prioritizes legal identity over semantic similarity. Parcel-specific retrieval is executed only after a successful parcel reference validation. Requests with missing, incomplete, invalid, or unresolved parcel references terminate before parcel lookup or regulatory retrieval, preventing unsupported parcel-specific reasoning.
Once the zoning row has been identified, a second retrieval step gathers supporting regulatory context. The query is enriched with parcel metadata, including zoning category, zoning index, and parcel purpose, while relevant parent rows from the zoning hierarchy are added to provide explanatory context. During answer synthesis, however, the exact category row always takes precedence over broader group- or territory-level information.

4.2 Deterministic Parameter Resolution and Planning Capacity

Once an exact zoning row has been selected, planning parameters are extracted from structured metadata rather than free-text parsing. Maximum FAR, maximum density, minimum green area, and maximum cornice height are normalized into scalar values whenever available. When zoning parameters are expressed as ranges, deterministic calculations are performed only if the required value is explicitly specified in the user query; otherwise, the corresponding parameter remains unresolved, and the calculation path is blocked.

When parcel area and zoning parameters are available, the workflow derives the main planning-capacity quantities. Let A_p denote parcel area, $\rho_{\max}$ the maximum density percentage, $\kappa_{\max}$ the maximum FAR, and $\gamma_{\min}$ the minimum green percentage. The primary planning quantities are

$$\begin{aligned} A_{\text{foot,max}} &= A_p \cdot \frac{\rho_{\max}}{100}, \\ A_{\text{gfa,max}} &= A_p \cdot \kappa_{\max}, \\ A_{\text{green,min}} &= A_p \cdot \frac{\gamma_{\min}}{100}. \end{aligned} \tag{1}$$

If existing development is available, the remaining parcel capacity is computed as

$$\begin{aligned} A_{\text{foot,rem}} &= A_{\text{foot,max}} - A_{\text{foot,exist}}, \\ A_{\text{gfa,rem}} &= A_{\text{gfa,max}} - A_{\text{gfa,exist}}. \end{aligned} \tag{2}$$

For proposal-oriented questions, feasibility is evaluated through a deterministic gate rather than free-form model reasoning:

$$\text{Feasible}(p) = \begin{cases} 1, & \text{if } A_{\text{foot,prop}} \le A_{\text{foot,rem}}, \\ & \quad A_{\text{gfa,prop}} \le A_{\text{gfa,rem}}, \\ & \quad h_{\text{prop}} \le h_{\max}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In practice, this feasibility is applied only when all the required planning parameters are available and sufficiently precise to support deterministic evaluation.

4.3 Parcel Side Normalization and Setback Assignment

Some parcel-specific questions depend not only on zoning parameters but also on the interpretation of parcel sides. Because cadastral polygons often contain fragmented boundary segments, the workflow first normalizes raw geometry into a smaller set of planning-oriented parcel sides before assigning setback rules.Neighbouring boundary segments are compared using a circular angular-difference function as follows:

$$\Delta(\beta_a, \beta_b) = \min\left(|\beta_a - \beta_b|,\ 180^\circ - |\beta_a - \beta_b|\right), \tag{4}$$

which avoids discontinuities caused by angular wrap-around and enables adjacent micro-fragments with similar orientation to be merged into a single planning side.

Street-frontage detection further depends on correctly constructing outward probe points around each normalized boundary segment. If $\alpha(s)$ denotes the segment azimuth returned by the spatial database, measured clockwise from north, then a probe of radius ρ is computed as

$$\Delta x = \rho \sin \alpha(s), \qquad \Delta y = \rho \cos \alpha(s). \tag{5}$$

Using the correct north-referenced convention is essential for reliable street-frontage detection and subsequent setback assignment. Incorrect directional assumptions can rotate the probe direction and lead to erroneous frontage identification.

Once normalized parcel sides have been established, setbacks are assigned according to side role and street class. For street-facing sides, the rule set is

$$\text{setback}_{\text{street}}(s) = \begin{cases} 15\text{ m}, & \text{frc}(s) = 1, \\ 10\text{ m}, & \text{frc}(s) = 2, \\ 3\text{ m}, & \text{frc}(s) \ge 3. \end{cases} \tag{6}$$

Neighbour-facing sides receive a default setback of 3 m, while the side identified as opposite the street frontage receives a 5 m setback. By separating geometric preprocessing from rule assignment, the workflow applies setback regulations deterministically rather than inferring them directly from text.

4.4 State-Aware Answer Synthesis

The final answer generation stage is route-aware rather than universal. The workflow uses distinct prompt families for general regulatory questions, parcel-specific contextual responses, and calculation- or feasibility-oriented answers. This keeps the prompting strategy aligned with the evidence available in the corresponding workflow state rather than relying on a single prompt for all question types. General regulatory and zoning questions use a concise general-answer prompt. Parcel-specific questions without deterministic calculations use a dedicated parcel-context prompt, while proposal and feasibility questions use a calculation-oriented prompt that receives precomputed planning quantities and validation results. Consequently, prompt selection depends not only on the question type but also on the evidence assembled during workflow execution.

Rather than free-form summaries, prompt inputs are constructed from structured workflow state. Depending on the route, the model receives explicitly separated blocks containing the user question, parcel metadata, building context, zoning parameters, legal evidence, setback information, deterministic calculations, and proposal-validation results. A deterministic answer mode (e.g., single-value, feasibility, setback, legal-reference, or lookup response) is also injected, removing the need for the language model to infer the response style.

For parcel-specific zoning questions, the workflow enforces hierarchy-aware final answer generation. When both an exact zoning-category row and broader parent rows are available, the category row always takes precedence, while higher levels are used only as explanatory context. This prevents unintended generalization from broader regulatory categories. The LLM is therefore used primarily for synthesis and verbalization rather than reasoning or computation. Prompts suppress internal workflow details, enforce concise Bulgarian responses, and use conservative decoding settings to maximize reproducibility. All answer routes currently use the same lightweight

reasoning model (`mistral-small3.1:latest`), while decoding parameters are adjusted slightly according to the amount of structured evidence provided. The configuration ranges are summarized in Table 2.

Route	Temp.	Top-p	Tokens	Context
General	0.05	0.75	220–250	12 288
Parcel-specific	0.05	0.75	220–250	12 288
Proposal	0.10	0.70	220–260	16 384

Table 2. Answer-generation parameters for the three response routes.

Low-temperature decoding, moderate top-p, and bounded output length prioritize reproducibility over linguistic variation and keep responses concise and evidence-grounded. The calculation route uses a slightly larger context window to accommodate additional structured state while preserving the same conservative generation strategy.

5. Benchmark Dataset and Evaluation Design

5.1 Benchmark Construction

The benchmark was designed as a reviewed, architecture-aware evaluation workflow rather than as a static question–answer dataset. The process begins with compact seed cases defining the question family, query type, parcel identifiers (where applicable), user question, and reviewer notes. From these seeds, the workflow generates structured evidence bundles containing parcel and building context, simplified parcel-side information, setback context, calculation references, and candidate responses for subsequent review.

The reviewed benchmark contains 40 approved cases. For each seed case, the workflow preserves the assembled evidence bundle together with the route information, retrieval traces, database context, calculation references, and runtime diagnostics. Candidate cases are imported into a review table, exposing both the generated answer and its supporting execution state, allowing reviewers to validate not only the final response but also the underlying workflow behaviour.

Human review remains an explicit methodological step. Reviewers correct expected outputs where necessary, reject weak or misleading candidates, and approve only validated cases for inclusion in the final benchmark. As a result, the benchmark evaluates not only answer quality but also route selection, parcel lookup, retrieval, deterministic calculations, and execution traceability, making it architecture-aware rather than a conventional one question–answer dataset. Parcel-reference issue cases are treated as a dedicated benchmark family. These include missing parcel references, incomplete parcel information, invalid identifier formats, and syntactically valid identifiers that return no parcel in the database. Including these cases ensures that bounded workflow failures are explicitly evaluated before unsupported parcel reasoning occurs.

Table 3 summarizes the composition of the reviewed benchmark across the principal workflow routes and evaluated capabilities. The benchmark families mirror the principal workflow routes rather than merely maximizing topical diversity. Together, they evaluate retrieval, parcel lookup, deterministic calculations, proposal validation, geometry-based reasoning, bounded failure modes, and out-of-domain behaviour, ensuring that the benchmark reflects the assistant's architecture rather than only the surface diversity of user questions.

Question family	N	Evaluation focus
General retrieval	6	Default legal retrieval; answer relevance
Legal-reference retrieval	1	Retrieval using explicit legal locators
Parcel zoning overview	6	Zoning lookup; parent-context expansion
Parcel lookup	2	Structured parcel lookup only
Parcel calculation	7	Planning-parameter extraction; deterministic calculations
Parcel proposal	6	Combined zoning reasoning; feasibility assessment
Parcel setback	2	Frontage interpretation; setback assignment
Parcel reference handling	6	Missing/invalid references; bounded failures
Out-of-domain	4	Routing discipline; bounded refusal

Table 3. Composition of the reviewed 40-case benchmark by evaluation family.

5.2 Route-Aware Metric Design

The evaluation framework combines deterministic and judge-based metrics. Deterministic evaluation provides the primary pass/fail mechanism for route selection, parcel lookup, and planning calculations, while judge-based metrics provide complementary graded assessment of answer quality, evidence use, and faithfulness where exact lexical matching is insufficient.

Parcel-reference issue cases require a dedicated evaluation policy. When the workflow correctly identifies missing, incomplete, invalid, or not-found parcel references, the expected outcome is a bounded deterministic response rather than evidence aggregation. The retrieval-context, combined-evidence, and answer-faithfulness metrics are not applied to this question family. Instead, evaluation focuses on route correctness, parcel-reference classification, and fixed-response matching. The same principle applies to out-of-domain questions, which are evaluated primarily as routing and bounded-behaviour cases.

An important methodological choice is that answer support is evaluated against the complete evidence bundle rather than retrieved text alone. Let a_q denote the final answer for query q and b_q the evidence bundle available during final answer generation. Architecture-aware faithfulness is therefore defined as

$$\mathrm{Faithfulness}(q) = f_{\mathrm{faith}}(a_q, b_q) \in [0, 1]. \tag{7}$$

The evidence bundle may contain retrieved legal passages, parcel lookup results, zoning parameters, deterministic calculations, proposal-validation state, or geometry-derived setback information, depending on the workflow route. Evaluating faithfulness against this combined evidence is more appropriate than comparing answers only with the retrieved text because many parcel-specific responses depend on a structured intermediate state. The same principle motivates the Combined Evidence Recall (CER):

$$\mathrm{CER}(q) = f_{\mathrm{cer}}(u_q, b_q, y_q^*) \in [0, 1], \tag{8}$$

where u_q is the user question and y_q^* is the expected answer representation. CER measures whether the complete evidence

bundle contains the information required to generate the expected answer, regardless of whether that information originates from free-text retrieval or structured workflow state.

Figure 4 summarizes the benchmark construction and evaluation workflow, from reviewed benchmark generation to execution tracing and complementary metric evaluation. The proposed route-aware evaluation avoids applying identical metrics to fundamentally different workflow routes. Metric applicability depends on the executed reasoning path, ensuring that retrieval, deterministic calculations, bounded failures, and out-of-domain behaviour are evaluated using criteria appropriate to each task.

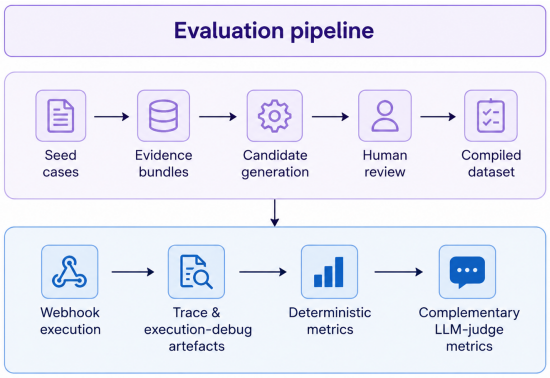

Figure 4. Simplified benchmark and evaluation workflow.

6. Results

Table 4 summarizes the aggregate performance on the reviewed 40-case benchmark. The evaluation combines deterministic metrics for routing, lookup, and structured-state correctness with complementary LLM-as-a-judge metrics assessing answer quality, grounding, and evidence use. This reflects the hybrid architecture of the assistant, where some properties require strict deterministic validation while others benefit from graded assessment.

Layer	Metric	Mean	Scored
Deterministic	Pass rate	1.000	40/40
	Routing	1.000	40/40
	Calculation	1.000	13/13
	Parcel-reference match	1.000	6/6
	Out-of-domain match	1.000	4/4
LLM-as-a-judge	Answer correctness	0.873	40
	Answer relevancy	0.828	40
	Faithfulness	0.931	40
	Combined evidence recall	0.858	30
	Calculation alignment	0.632	19
	Calculation interpretation	0.571	21

Table 4. Aggregate benchmark results by evaluation metric.

The deterministic evaluation achieved perfect performance across all applicable cases. Routing was correct for all 40 benchmark questions, deterministic planning calculations matched the expected outputs in all 13 applicable cases, and bounded-response behaviour was correct for both parcel-reference issue cases and out-of-domain questions. These results demonstrate that the workflow consistently executes the intended route-specific logic and deterministic reasoning. The judge-based evaluation provides a more differentiated assessment of the generated answers. Answer correctness reached 0.873, answer relevancy 0.828, and faithfulness 0.931. Because faithfulness is evaluated against the complete structured evidence bundle rather than retrieved text alone, it primarily measures whether answers remain grounded in the assembled workflow evidence.

Performance varies across the benchmark families summarized in Table 3. Parcel-calculation cases achieved the strongest results, with answer correctness of 0.964 and faithfulness of 1.000, reflecting the benefits of deterministic planning calculations. Proposal-feasibility questions achieve an answer correctness of 0.690 and faithfulness of 0.800 despite perfect deterministic calculation performance. Parcel-zoning overview questions occupy an intermediate position, suggesting that the remaining limitations are mainly related to answer synthesis rather than retrieval or routing. The reviewed setback cases were answered correctly, confirming that the geometry-based preprocessing pipeline functions as intended.

6.1 Error Analysis

The reviewed benchmark indicates that the principal remaining limitation lies in final-answer synthesis rather than in deterministic workflow execution. Proposal-feasibility questions illustrate this most clearly. In several cases, the structured proposal checks and deterministic conclusions were correct, yet the generated answer remained overly cautious or insufficiently explicit about the governing planning constraint. This explains why proposal cases achieved perfect deterministic performance while receiving lower judge-based scores for answer correctness and calculation interpretation.

Parcel-zoning overview questions reveal a second source of error related to hierarchy-aware synthesis. Although the correct zoning category was retrieved, the final response occasionally generalized toward broader parent categories, reducing the specificity of parcel-level explanations. This suggests that future improvements should focus on strengthening hierarchy-aware answer generation rather than retrieval itself.

A third source of variation arises from the judge-based evaluation. LLM judges occasionally assign lower scores to concise but factually correct responses that omit secondary explanatory details. This effect is most evident in proposal-oriented questions and should be interpreted alongside the deterministic evaluation rather than as evidence of deficiencies in routing, retrieval, or planning calculations.

In conclusion, the observed errors are concentrated in answer verbalization rather than in deterministic routing, retrieval, or planning calculations, indicating that future improvements can focus primarily on the synthesis layer while preserving the existing workflow architecture.

7. Discussion, Limitations, and Future Work

The proposed assistant should be viewed as a workflow-oriented system for first-pass urban planning analysis. Its main contribution is not unrestricted question answering, but the explicit integration of route-aware execution, structured spatial data, regulatory retrieval, deterministic planning calculations, and evidence-grounded answer synthesis. This design makes the workflow transparent, inspectable, and directly aligned with

the architecture-aware evaluation framework presented in this paper.

The benchmark results demonstrate that the proposed architecture provides a robust foundation for parcel-level planning assistance. Deterministic routing, parcel lookup, and planning calculations performed consistently across the reviewed benchmark, indicating that the principal remaining challenge lies in communicating structured evidence rather than producing it. The observed limitations are therefore concentrated in answer synthesis, particularly for proposal-feasibility questions and parcel-zoning explanations, where responses would benefit from clearer articulation of governing constraints and more consistent preservation of zoning hierarchy.

The current workflow intentionally adopts a conservative strategy for deterministic calculations. Calculations are executed only when the required planning parameters are available with sufficient precision, preventing the assistant from producing unsupported quantitative conclusions. While this limits the scope of some queries, it increases the reliability and interpretability of the generated planning advice. The evaluation methodology also highlights opportunities for refinement. Faithfulness is assessed against the complete route-aware evidence bundle rather than the retrieved text alone, reflecting the assistant's hybrid nature. Future work should further refine judge prompts and evaluation criteria so that concise, evidence-grounded answers are distinguished more consistently from genuinely incomplete responses.

Several research directions follow naturally from the present work. First, the benchmark will be expanded with more diverse planning scenarios and more complex route combinations to improve evaluation coverage. Second, the assistant will be extended with richer regulatory knowledge representation, additional planning documents, and more advanced retrieval strategies, including hybrid lexical-semantic retrieval and graph-based regulatory modeling. Third, future versions will incorporate memory-enabled multi-turn interaction together with improved route detection and more capable language models to strengthen parcel-specific explanation and proposal evaluation. Finally, the workflow will be extended beyond parcel-level assessment toward building-aware reasoning, enabling more detailed analysis of complex urban development scenarios.

Although developed for the Bulgarian context, the proposed methodology is not inherently jurisdiction-specific. The combination of structured data preparation, route-specialized workflow execution, and architecture-aware evaluation provides a general design pattern for AI assistants operating in regulation-intensive domains. The Sofia case study, therefore, serves not only as an application example, but also as a validation of a broader methodology for trustworthy geo-legal decision support.

8. Conclusion

This paper presented an agentic urban planning assistant for parcel-level spatial and regulatory evaluation. The central contribution is a route-aware workflow that combines structured spatial and regulatory data, specialized retrieval, deterministic planning calculations, and evidence-grounded answer synthesis. The results demonstrate that parcel-level planning assistance cannot be reduced to generic retrieval-augmented generation alone, but instead requires explicit orchestration of deterministic and language-based reasoning.

Additional contribution is an architecture-aware evaluation framework that combines reviewed benchmark construction, route-specific metric policies, deterministic assertions, complementary LLM-as-a-judge evaluation, and execution-level observability. The reviewed benchmark demonstrates that the proposed workflow achieves robust deterministic routing, parcel lookup, and planning calculations while identifying answer synthesis as the principal remaining area for improvement.

Beyond the Sofia case study, the proposed methodology provides a general framework for trustworthy AI assistance in regulation-intensive domains. By combining route-aware workflow design with transparent evaluation, it offers a foundation for future geo-legal assistants that require traceable, deterministic, and evidence-grounded decision support.

Acknowledgement

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155 and the programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, the GENeCITY project, funded by the Horizon Europe programme under grant agreement no. 101270173.

References

Bruzzone, A., Giovannetti, A., Genta, G., Cefaliello, D., 2024. Generative AI and Retrieval-Augmented Generation (RAG) in an Agent-Based Simulation Framework for Urban Planning. *Proceedings of the 23rd International Conference on Modelling and Applied Simulation*, CAL-TEK srl.

Cao, Z., Shen, L., Xu, X., Guo, Y., Pan, B., Bao, H., 2026. A Retrieval-Augmented Generation (RAG) Based Framework for Evaluating Urban Low-Carbon Governance and Its Implications for Sustainable Development. *Sustainability*, 18(4), 2143.

Chen, T., Miao, S., Cai, S., Liu, H., Mu, Y., Yu, K., Deng, H., Jin, X., Jiang, Q., Jiang, H., Wang, P., Ma, D., Zahir, A., 2026. GeoRAG: A Geographic Retrieval Augmented Generation Framework Based on Urban Spatio-Temporal Knowledge Graph. X. Ding, Y. Dong (eds), *Computational Linguistics and Natural Language Processing*, 2775, Springer Nature Singapore, Singapore, 130–140.

Kalyuzhnaya, A., Mityagin, S., Lutsenko, E., Getmanov, A., Aksenkin, Y., Fatkhiev, K., Fedorin, K., Nikitin, N. O., Chichkova, N., Vorona, V., Boukhanovsky, A., 2025. LLM Agents for Smart City Management: Enhancing Decision Support Through Multi-Agent AI Systems. *Smart Cities*, 8(1), 19.

Ku, H.-C., Zhou, C.-H., Tsai, P.-W., Lee, S.-H., 2025. Application of the Retrieval-Augmented Generation Method in Consultation Systems for Smart Cities. *2025 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, IEEE, Kaohsiung, Taiwan, 783–784.

Li, Z., Xia, L., Tang, J., Xu, Y., Shi, L., Xia, L., Yin, D., Huang, C., 2024. UrbanGPT: Spatio-Temporal Large Language

Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, Barcelona Spain, 5351–5362.

Liu, R., Zhe, T., Peng, Z.-R., Catbas, N., Ye, X., Wang, D., Fu, Y., 2025a. Towards Urban Planing AI Agent in the Age of Agentic AI. arXiv.

Liu, Y., Yin, Z., Wang, J., Li, K., 2025b. TransRAG: A Novel Hybrid Retrieval Augmented Generation Framework for Public Transportation Application. *2025 37th Chinese Control and Decision Conference (CCDC)*, IEEE, Xiamen, China, 66–71.

Na Takuatung, S., Bussracumpakorn, C., 2026. A Comparative RAG Agent vs. Base LLM for Community-Based Tourism Recommendations: A Pre-Deployment Validation Study. *2026 International Conference on AI Innovations and Industry (ICAIII)*, IEEE, Jeddah, Saudi Arabia, 1–6.

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.

Salazar-Miranda, A., Talen, E., 2025. Zoning in American Cities: Are Reforms Making a Difference? An AI-based Analysis. arXiv.

Wang, C., Yan, X., Dai, Y., Wang, Z., Xu, S., 2026. From Image Generation to Infrastructure Design: A Multi-agent Pipeline for Street Design Generation.

Wang, D., Fu, Y., Wang, P., Huang, B., Lu, C.-T., 2020. Reimagining City Configuration: Automated Urban Planning via Adversarial Learning. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, ACM, Seattle WA USA, 497–506.

Xie, X., Herrera, M., Tang, X.-Y., James, P., Kassem, M., 2026. Trustworthy Urban Digital Twin: A RAG-based Architecture for Integrating Verifiable Knowledge. *IET Conference Proceedings*, 2026(1), 43–47.

Xu, H., Yuan, J., Zhou, A., Xu, G., Li, W., Ban, X., Ye, X., 2024. GenAI-powered Multi-Agent Paradigm for Smart Urban Mobility: Opportunities and Challenges for Integrating Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) with Intelligent Transportation Systems. arXiv.

Zhou, Z., Lin, Y., Jin, D., Li, Y., 2024. Large Language Model for Participatory Urban Planning. arXiv.