Spatial RAG for Big Data–Driven Urban Itinerary Recommendation

Maddalena Amendola[1,†], Chiara Pugliese [1,†], Raffaele Perego[2] and Chiara Renso[2]

[1]IIT-CNR, Pisa,Italy, maddalena.amendola@iit.cnr.it, chiara.pugliese@iit.cnr.it
[2]ISTI-CNR,Pisa, Italy, raffaele.perego@isti.cnr.it, chiara.renso@isti.cnr.it,

Abstract

Urban mobility and tourism rely on heterogeneous Big Data, yet traditional recommender systems often rely on static shortest-path models that ignore complex, dynamic urban indicators. While LLMs offer reasoning capabilities that can mitigate these issues, they lack factual accuracy and are prone to spatial hallucinations. We present UrbanRAG, a Spatial Retrieval-Augmented Generation (RAG) framework that fuses LLMs with neural information retrieval to leverage large-scale urban Big Data. UrbanRAG enables natural-language interaction to generate personalized itineraries by integrating worldwide map data, environmental indicators, and other semantic information. Unlike localized databases, our open-world architecture ensures scalability and portability. Preliminary empirical results demonstrate that UrbanRAG significantly improves spatial grounding and accuracy over closed-book LLMs. By anchoring generative models in real-world spatial data, our framework provides a robust, adaptive solution for data-intensive urban services.

1. Introduction

Urban mobility and tourism increasingly rely on large-scale, heterogeneous data sources describing cities, human activities, and environmental conditions. In this context, Big Data technologies enable the integration and analysis of massive volumes of spatial, contextual, and user-generated information, opening new opportunities for data-driven, personalized services. While several studies have explored data-intensive systems for urban itinerary recommendation [1], traditional approaches often simplify the problem to static shortest-path optimization, failing to fully exploit the multidimensional nature of available urban data and user preferences. The semantic integration of diverse geographical Big Data remains an open challenge, with no universally accepted solution having been found to date. While Large Language Models (LLMs) are widely used nowadays for complex reasoning, they exhibit limitations in factual accuracy and spatial retrieval when used in isolation [2, 3]. Retrieval-Augmented Generation (RAG) [4] addresses this by combining LLM generative capabilities with external data retrieval, reducing hallucinations and ensuring factual consistency. Inspired by [3], we argue that RAG-based architectures provide a solid foundation for spatially enabled applications that rely on Big Data. By coupling LLMs with large-scale spatial, environmental, and contextual datasets, Spatial RAG systems can retrieve geographically grounded information, integrate diverse urban indicators, and generate personalized recommendations.

In this work, we present UrbanRAG, a Spatial RAG framework in which an LLM orchestrates spatial reasoning and information retrieval (IR) through a conversational interface to support adaptive, user-centric urban itinerary recommendation. Users express preferences and spatial constraints in natural language, enabling dynamic retrieval of route semantics and geographic features. Unlike approaches limited to local spatial databases, UrbanRAG leverages open, worldwide map data and integrates heterogeneous urban Big Data. UrbanRAG represents a general RAG-based methodology for integrating spatial and contextual information, moving beyond static indicator–based models toward adaptive reasoning driven by evolving data and user feedback. To our knowledge, it is the first framework to jointly support conversational interaction, spatial reasoning, and neural IR over urban Big Data. Our prior system, RAGTrip [5], is a specific instantiation of this paradigm, while UrbanRAG generalizes

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
[†]These authors contributed equally.

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-09

CEUR-WS.org/Vol-4192/BMDA-paper4.pdf

the integration of Spatial and IR components and includes a comprehensive empirical evaluation. Reproducible experiments demonstrate that RAG-based spatial integration significantly improves accuracy, outperforming closed-book LLMs prone to spatial hallucinations.

In summary, this paper makes the following contributions: (1) a spatial RAG system to fuse large-scale spatial data, environmental indicators, and unstructured textual corpora; (2) an open-world, scalable architecture that avoids geographically constrained databases; (3) an empirical evaluation through a walkability-based itinerary recommendation showing how spatial RAG mitigates hallucinations and improves spatial grounding.

2. The UrbanRAG framework

UrbanRAG interacts with users through a conversational interface, complemented with a backend which performs the actual itinerary computation. The framework comprises three main components, depicted in Figure 1: the Query Understanding and Answer Generation (QUAG) component, the Spatial component, and the IR component. We detail the modules' functionalities below.

2.1. Query Understanding and Answer Generation

This component encapsulates an LLM and manages conversational interactions with users. Upon receiving an utterance, it determines whether the user's information need pertains to: (i) a new itinerary suggestion taking into account semantic indicators and user preferences, or (ii) general information about attractions or points of interest (POIs) related to a previously suggested itinerary. In the first case, the query is routed to the Spatial component of UrbanRAG, which generates an itinerary (if any) between the specified points, constrained by auxiliary information that may be of interest to the user. All other queries not involving itinerary requests are instead forwarded to the IR component.

In both scenarios, the information retrieved by the Spatial or IR components is leveraged by QUAG to augment the LLM's generation process, producing the final response grounded in the retrieved content. When the Spatial component is involved, the response includes details of the suggested itinerary, emphasizing the route's semantic features and the POIs encountered. In the case of a general information request, the *top-k* results retrieved by the IR component from a knowledge repository are leveraged by QUAG to generate a appropriate response, improving factual accuracy and reducing hallucinations.

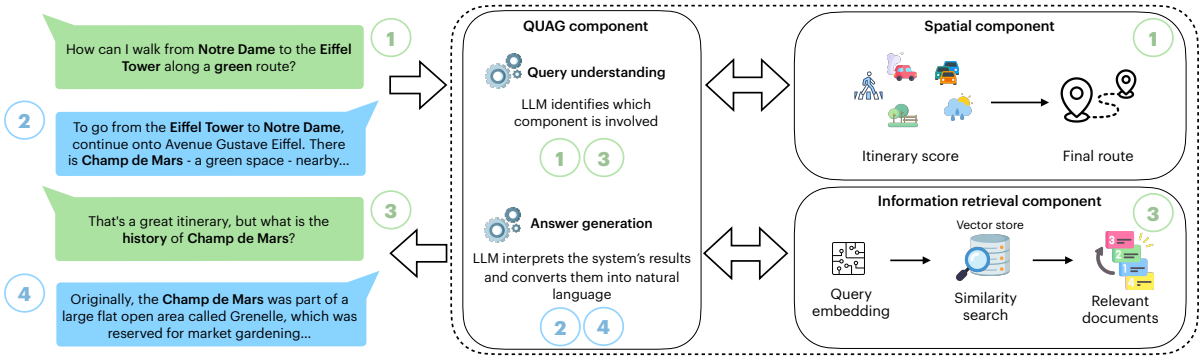

Figure 1: Overview of the UrbanRAG framework.

2.2. Spatial component

Once the QUAG component determines that a user has queried for an itinerary, it routes the request to the spatial data fusion reasoning engine. This component identifies the optimal itinerary based on users' requests between the specified origin and destination. Here, we focus on *domain-specific spatial*

indicators to identify optimal paths. These indicators are pluggable to the model and can be generalized to other domains (such as safety, green, and equity). They are combined into a generic multi-criteria spatial scoring to generate the most appropriate itinerary based on the users' requests. Examples of these multi-criteria scoring methods include combining features such as road network characteristics (e.g., pedestrian paths), environmental data (e.g., air pollution), the presence of green areas, shops, and accessibility.

We quantify the overall domain-specific indicator score of each candidate route as follows. First, for each segment along a route, we count the occurrences of indicators, capping each segment's contribution at a maximum threshold τ to mitigate the impact of outliers. Second, for each indicator, we compute the average capped count per segment, denoted c_i, by dividing the total capped count by the number of segments. Third, we assign a user-defined weight w_i to each indicator to reflect its relative importance, with the constraint $\sum_i w_i = 1$. In the absence of user input, uniform weighting is applied. The final score for the route is calculated as $WS = \frac{\sum_i w_i \, c_i}{\tau}$. In addition to domain-specific indicators, the Spatial component enriches the route based on user preferences or contextual information. When such preferences are explicitly expressed in the query, the QUAG component identifies them and forwards the relevant parameters to this component. Examples of such preferences could include requesting a vegan restaurant along the path or visiting a kid-friendly museum before arrival. To incorporate these preferences, we generate a buffer around each alternative route and perform a spatial join to associate additional POIs or relevant features requested by the user. If no explicit preferences are provided, we enrich routes with general tourist information to ensure the itinerary remains informative and engaging.

2.3. Information Retrieval component

The IR component integrates a neural indexing and search system. Offline, documents from the knowledge base are encoded into dense vectors within a multidimensional latent space and stored in a vector index, enabling efficient similarity-based retrieval. At query time, each user query received by QUAG is similarly encoded into the same latent space. The resulting query representation is then compared to the indexed document vectors using an approximate nearest-neighbor search algorithm. The top-k closest vectors retrieved from the index are considered the relevant context. The associated documents are returned to QUAG, which uses them to generate a grounded and contextually informed response.

3. UrbanRAG assessment

In this section, we describe the implementation of the three components of UrbanRAG, outline the experimental settings, and discuss the results along with the insights gained.

QUAG Component. The QUAG component is implemented in Python and manages the RAG-based conversational interface, which encapsulates the Llama 3.1 8B[1] model as LLM for query classification and retrieval-augmented answer generation.

Spatial Component. After receiving from QUAG an itinerary request including the start and end points, we use the `Nominatim` API[2] to obtain the corresponding latitude and longitude. Then, we generate three alternative routes using the GraphHopper API[3], which queries the OpenStreetMap road network. The contextual features are retrieved through appropriate API calls: for example, the air quality index is retrieved by using OpenWeatherMap's API[4], instead, the POIs are retrieved with the `OSMnx` library by filtering OpenStreetMap data using relevant tags, such as land use, natural, footway, wheelchair, and tourism. The output returned to QUAG is a JSON file containing information about

[1] https://huggingface.co/meta-llama/Llama-3.1-8B
[2] https://nominatim.org/
[3] https://www.graphhopper.com/
[4] https://openweathermap.org/api

the route with the highest score based on the user itinerary request, i.e., the routing instructions, the itinerary score and indicators, and the list of POIs associated with each segment, together with their category.

IR Component. For information queries, we use the TREC Conversational Assistance Track (CAsT) 2019 and 2020 collections [6, 7], comprising TREC CAR, MS MARCO [8], and the Washington Post corpus, for a total of 38,636,520 passages. To retrieve relevant passages in response to queries, we leverage the FAISS vector search library [9, 10] and the `Snowflake`[5] bi-encoder, built on XLM-R Large and fine-tuned for retrieval tasks [11]. Passages are encoded and indexed offline as 1024-dimensional dense embeddings. At query time, the system encodes the incoming query, computes cosine similarity against the pre-computed embeddings, and returns the top-3 most relevant passages to QUAG for answer generation.

UrbanRAG Dataset. For evaluation, we focus on *the most walkable itinerary* in the city of Paris. However, the approach is generic and can be easily adapted to other itinerary indicators, such as less-polluted, quieter, more culturally rich, or greener options. We constructed a custom dataset comprising realistic walking and information-seeking queries in the city of Paris. The dataset includes 10 distinct spatial requests, each paired with 3 follow-up information queries regarding the route or nearby landmarks, for a total of 40 user queries. This design simulates a typical user interaction: a person first requests a walkable route and then engages in conversation about the places encountered along the way.

Reproducibility. The dataset and the LLM instructions used are already available in our GitHub repository[6].

Evaluation. We compare the accuracy of the answers generated for the queries in our dataset by two system configurations: (i) **UrbanRAG**, our open-book framework where spatial queries are answered based on route and environmental indicators by the spatial component, and information queries are grounded using the top-k passages retrieved by our IR component; (ii) **LLM-ClosedBook (LLM-CB)**, a baseline configuration where the LLM model (Llama 3.1 8B) is used in isolation without any route enrichment or external retrieval augmentation.

3.1. Results and discussion

In this section, we present the results of our evaluation, focusing separately on the two types of user interactions addressed by UrbanRAG: *Spatial* and *Information* requests.

Specifically, Table 1 presents our preliminary evaluation using the queries of UrbanRAG dataset. We evaluate 10 spatial requests by assessing the LLM's ability to generate coherent and walkable routes based on user preferences. For the 30 Information requests, we assess the model's ability to provide accurate and appropriate answers to general-purpose queries about urban entities encountered along the route. We further analyze whether UrbanRAG mitigates the limitations of the closed-book LLM (LLM-CB) in both spatial and informational tasks by leveraging contextual and geographic knowledge. Responses are labeled as correct, partially correct (minor uncertainties), or incorrect (completely wrong).

Table 1
Summary of the evaluation results

System	Query type	Correct	Partially correct	Incorrect
LLM-CB	Spatial	0	0	10
	Information	12	11	7
UrbanRAG	Spatial	4	6	0
	Information	20	5	5

Query understanding. In our experiments, the integrated LLM allowed the QUAG component to correctly classify all 40 queries and route them to the appropriate spatial or IR component.

[5] https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0
[6] https://github.com/chiarap2/urbanRAG

Spatial requests. To evaluate the effectiveness of the proposed Spatial RAG mechanism, we analyzed responses to 10 route-based queries in our dataset. A route is considered correct if it leads the user from the specified origin to the destination in a continuous way, taking into account walkability indicators and any user-defined preferences. LLM-CB failed all 10 queries. Its responses consistently contained hallucinations, such as suggesting directions that exhibited significant jumps from the intended path (ranging from 1.7 km to 8.6 km), looping instructions, and poor spatial awareness (e.g., confusion between left and right). Furthermore, it frequently recommended POIs far from the actual route, such as in the third spatial query of the dataset, where LLM-CB suggested visiting *Café de la Paix*, which lies 3.9 km from the *Jardin des Plantes* destination. This example is shown in Figure 2. By contrast,

Figure 2: LLM-CB and UrbanRAG routes for the third spatial query of the dataset.

UrbanRAG returned 4 fully correct routes and 6 partially correct ones. Partial correctness was defined by minor omissions in navigation steps. To achieve these results, we experimented with various instruction formulations. Less structured prompts produced responses that were more fluent but less accurate, whereas the more schematic prompt resulted in higher accuracy but reduced textual fluency. Importantly, UrbanRAG did not produce hallucinations. In most of the partially correct responses, the missing steps were duplicates of earlier instructions (e.g., repeated turns or identical POIs). In terms of walkability, UrbanRAG accurately identified both of the poorly walkable routes in the dataset (the fourth and the fifth), incorporating user preferences into its assessment. LLM-CB, on the other hand, correctly flagged only one of these, and only due to an overestimation caused by an erroneous 10-km route. For the other, it recommended public transportation but directed users to a station roughly 5 km away from the intended destination. Notably, across all LLM-CB responses, the initial and final instructions were typically aligned with the start and end points, while the intermediate steps exhibited substantial disorientation.

Information requests. To evaluate the effectiveness of the IR-based RAG mechanism in UrbanRAG, we analyzed the system's responses to the 30 information queries included in our dataset. Each response was manually labeled as correct, incorrect, or partially correct (imprecise). A response was deemed *partially correct* when it conveyed the correct general information but included factual inaccuracies or lacked specificity. For instance, a typical partially correct answer occurred when responding to the question *"What are the most important hotels in Paris?"*. In this case, the model listed several relevant hotels but also included the *Hotel du Petit Bourbon*, a historical building demolished in the 17th century. While the answer captures the intended topic (notable hotels), it failed to distinguish between current and historical relevance.

The results achieved clearly highlight the benefit of RAG: UrbanRAG returned 20 correct answers, 5 partially correct answers, and 5 incorrect answers. Notably, in 3 of the incorrect answers, the system failed to retrieve relevant information from the indexed collection, which prevented the LLM from

generating a grounded response. LLM-CB, by contrast, produced 12 correct answers, 11 partially correct, and 7 incorrect ones. These results confirm our study's premise: effective urban spatial reasoning requires grounding LLMs in large-scale, heterogeneous data. By integrating open map data and environmental indicators and contextual data via a unified Spatial RAG architecture, UrbanRAG successfully exploits large urban data to mitigate hallucinations. This ensures consistent, preference-aware itineraries that closed-book LLMs cannot produce without structured, real-world data.

4. Conclusions and future work

We introduce UrbanRAG, a spatially enhanced retrieval-augmented framework equipped with a conversational interface that recommends personalized urban itineraries by integrating large amounts of heterogeneous geographical data. Our experiments show that RAG significantly enhances factual accuracy and completeness. While UrbanRAG may falter when retrieval lacks sufficient context, it consistently outperforms closed-book LLMs, which are prone to spatial/factual hallucinations and exhibit limited contextual knowledge. Our findings highlight how LLMs alone struggle to generate coherent itineraries or to suggest urban elements in an exploratory context. They also underperform on general-purpose queries about less popular topics. These preliminary results open avenues for future work: (1) assessing the impact of different LLM model sizes and architectures on RAG performance; (2) enhancing spatial reasoning through richer geographic operations and the use of routing algorithms; and (3) studying how LLMs process structured route data (particularly their tendency to omit repeated instructions), potentially via improved spatial encoding in the RAG pipeline.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgments

This work was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), funded by the European Commission under the NextGeneration EU programme, by the MUSIT Project through the European Union's Horizon 2020 Research and Innovation program under MSCA GA no. 101182585, and by the CAMEO PRIN project (Research Grant no. 2022ZLL7MW) funded by the Italian Ministry of Education and Research (MUR), and by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] C. Panagiotakis, E. Daskalaki, H. Papadakis, P. Fragopoulou, An expectation-maximization framework for personalized itinerary recommendation with poi categories and must-see pois, ACM Transactions on Recommender Systems 3 (2024) 1–33.

[2] F. Li, D. C. Hogg, A. G. Cohn, Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark, in: Proc. AAAI Conf. on AI, volume 38, 2024, pp. 18500–18507.

[3] D. Yu, R. Bao, G. Mai, L. Zhao, Spatial-rag: Spatial retrieval augmented generation for real-world spatial reasoning questions, 2025. URL: https://arxiv.org/abs/2502.18470. arXiv:2502.18470.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[5] C. Pugliese, M. Amendola, R. Perego, C. Renso, A spatially-grounded conversational planner for personalized urban itineraries, in: Proc. 33rd ACM International Conference SIGSPATIAL, ACM, 2025, p. 840–843.
[6] J. Dalton, C. Xiong, V. Kumar, J. Callan, CAsT-19: A dataset for conversational information seeking, in: Proc.ACM SIGIR, ACM, 2020.
[7] J. Dalton, C. Xiong, J. Callan, CAsT 2020: The conversational assistance track overview, TREC'20, Virtual, 2020. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf.
[8] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, Ms marco: A human-generated machine reading comprehension dataset (2016).
[9] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.
[10] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). arXiv:2401.08281.
[11] P. Yu, L. Merrick, G. Nuti, D. Campos, Arctic-embed 2.0: Multilingual retrieval without compromise, arXiv:2412.04506 (2024).