PeRAG: Multi-Modal Perspective-Oriented Verbalization with RAG for Inclusive Decision Making

Muhammad Saad Amin[1,2,*], Horacio Jesús Jarquín Vásquez[1], Franco Sansonetti[1], Simona Lo Giudice[3], Valerio Basile[1] and Viviana Patti[1]

[1]Department of Computer Science, University of Turin, Italy
[2]Department of Electrical and Computer Engineering, Aarhus University, Denmark
[3]Dipartimento di Economia e Statistica "Cognetti de Martiis", University of Turin, Italy

Abstract

Urban policy makers require comprehensive insights into transportation issues and demographic distributions to design equitable and efficient infrastructure. However, analyzing multi-modal data (numeric and visual) while accounting for diverse perspectives remains challenging. To address this, we propose PeRAG, a novel pipeline combining multi-modal perspective-oriented verbalization with Retrieval-Augmented Generation (RAG). Our approach first converts numeric transportation/demographic data and population heatmaps into natural language descriptions using LLaMA, incorporating multiple policy-relevant perspectives. These verbalizations are then fed into the RAG system to generate context-aware, perspective-driven responses for urban planners. We demonstrate the effectiveness of PeRAG in generating actionable insights for transportation policy, bridging the gap between raw data and decision-making. Our experiments highlight the pipeline's ability to handle heterogeneous data modalities while adapting to diverse stakeholder viewpoints, offering a scalable solution for smart city analytics.

Keywords

Multi-modal Verbalization, Retrieval-Augmented Generation (RAG), Perspective-Aware NLP, Large Language Models (LLMs), Urban Transportation Analytics

1. Introduction

Urban policy makers face significant challenges in designing equitable transportation systems due to the complex interplay of demographic shifts, infrastructure constraints, and socio-economic disparities [1]. Raw data (e.g., transit logs, census metrics, heatmaps) is often siloed, requiring labor-intensive integration to derive insights [2, 3]. While NLP and computer vision techniques have been applied to urban analytics, they typically treat data modalities independently, ignoring the need for cross-modal reasoning (e.g., correlating heatmap patterns with numeric poverty indices) [4]. This limits their utility for policy decisions requiring holistic, interpretable inputs.

In recent years, advances in machine learning and NLP have enabled new forms of automated data interpretation, particularly in multimodal settings where information spans both structured and unstructured modalities [5]. Urban environments provide a rich case for multimodal reasoning: data can include numerical variables (e.g., population size, number of transport lines), visual artifacts (e.g., heatmaps of population density), and geographical descriptors (e.g., district boundaries). Integrating and interpreting these different modalities coherently is essential for supporting informed decision-making.

One of the emerging challenges in this context is perspective-aware verbalization, the task of transforming multimodal data into textual descriptions that reflect different analytical or stakeholder viewpoints [6]. For instance, the same urban dataset can be verbalized from a demographics perspective ("This area has a high population of elderly residents") or a transportation accessibility perspective ("This zone has limited coverage of public transport lines despite high population density"). Generating such targeted descriptions from numeric and image data requires models that understand not only the input modalities but also the intended angle of interpretation [7]. This introduces both linguistic complexity—in choosing appropriate vocabulary, structure, and focus—and reasoning complexity—in determining what information is salient for a given perspective.

These challenges compound when integrated into retrieval-augmented generation (RAG) pipelines. Traditional RAG frameworks are typically designed for text-based retrieval from large knowledge bases; extending them to operate over generated textual representations of multimodal data introduces new issues: retrieval is

---

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.
saad@ece.au.dk (M. S. Amin); horaciojesus.jarquinvasquez@unito.it (H. J. Jarquín Vásquez); franco.sansonetti@unito.it (F. Sansonetti); simona.logiudice@unito.it (S. Lo Giudice); valerio.basile@unito.it (V. Basile); viviana.patti@unito.it (V. Patti)
0000-0002-7002-9373 (M. S. Amin); 0000-0001-8110-6832 (V. Basile); 0000-0001-5991-370X (V. Patti)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

only as effective as the fidelity and perspective alignment of the verbalized input, and generation must remain factual, grounded, and contextually relevant [8]. Moreover, multimodal verbalizations are often more compact and abstract than traditional long-form documents, which poses difficulties in relevance ranking and context-aware generation.

In this work, we investigate the following core research questions:

1. How can multimodal data (numeric and visual) be verbalized in a perspective-aware manner to support policy-level interpretation?
2. What are the linguistic and functional trade-offs between zero-shot and few-shot verbalization approaches in this context?
3. Can a lightweight, locally-deployable RAG pipeline (PeRAG) effectively answer urban policy questions when built on top of such verbalizations?
4. How does the factuality and utility of such a system compare to general-purpose LLMs, especially in high-stakes policy scenarios?

To address these questions, we present PeRAG, a novel framework that combines multimodal data verbalization with a perspective-aware Retrieval-Augmented Generation pipeline. Our work is based on a custom dataset for the city of Turin, comprising over 7,000 examples across multiple years (2012–2019), including 31 features covering demographics, transportation, and traffic. We verbalize both numeric and heatmap data into English summaries across several perspectives (e.g., demographics-focused, transport lines-focused, temporal shifts), using *LLaMA-3.1-8B* for the verbalization of numeric data, and *LLaMA-3.2-11B-Vision* for the verbalization of heatmap data in zero-shot and few-shot settings. These verbalizations serve as the retrievable memory in a *Gemma-3-4B-IT*-powered RAG system, which supports question-answering on urban policy issues. All models are run locally to ensure data privacy and control.

Our key contributions are as follows:

- We introduce a multi-modal perspective-aware verbalization pipeline that generates textual summaries from numeric and image data for urban policy domains.
- We propose and implement PeRAG, a lightweight RAG-based QA framework grounded in multimodal verbalizations, optimized for locally-deployable urban analytics.
- We explore and analyze zero-shot vs. few-shot verbalization strategies in real-world settings, providing insight into generation fidelity and perspective alignment.
- We conduct human evaluation and qualitative analysis to assess factuality and relevance, and compare PeRAG outputs against general-purpose LLMs.

The rest of this paper is organized as follows: Section 2 reviews related work in multimodal NLP, verbalization, and RAG systems. Section 3 describes the methodology, including dataset details, verbalization techniques, and system architecture. Section 4 outlines our experimental setup. Section 5 presents results from verbalization and QA evaluations. Section 6 offers a detailed analysis and discussion. Section 7 concludes the paper and outlines directions for future work.

2. Related Work

Perspectivism in NLP is an emerging approach that emphasizes representing and reasoning with multiple, potentially divergent, viewpoints. Traditional NLP systems often adopt a mono-perspective stance, optimizing for a generalized “truth” or majority viewpoint. In contrast, recent work has called for more inclusive approaches that recognize and operationalize multiple coexisting viewpoints [9, 10].

In this context, Data Perspectivism [10] proposes that AI systems—especially in socially sensitive domains—should be capable of tailoring outputs to the values and expectations of distinct population segments. Though this paradigm has influenced tasks in Natural Language Understanding (NLU), its application in Natural Language Generation (NLG)—especially from heterogeneous data sources—is still under-explored. Our work addresses this gap by extending perspectivist reasoning to the generation of text from multimodal data, creating perspective-conditioned verbalizations that help communicate the same data through different analytical and social lenses.

Current multi-modal NLP approaches integrate structured and unstructured sources—such as tables, images, and text—but usually with the aim of generating a single canonical description (e.g., image captioning or data-to-text generation). What is largely missing is the ability to generate multiple alternative descriptions of the same input, each aligned with a distinct interpretive frame.

Our work situates itself uniquely at this intersection—producing diverse textual outputs from structured numeric and image-based urban data, each representing a different lens (e.g., accessibility for elderly residents, environmental concerns, transit network optimization). In this way, we operationalize perspectivism across modalities and offer diverse conditioned NLG from heterogeneous sources, a setting largely unexplored in current literature.

A major research avenue in knowledge-enhanced language modeling is Retrieval-Augmented Generation (RAG), in which a retriever module selects relevant textual passages from a knowledge base that are then fed into a generator to produce a grounded, informative response [8, 11]. This has been particularly effective in tasks like open-domain QA, summarization, and dialogue. Variants such as MuRAG [12] have explored incorporating multiple modalities into retrieval pipelines.

In our work, we adapt and extend the RAG architecture for perspective-aware generation by populating the retrieval index with natural language verbalizations that encode distinct viewpoints over the same input data. Unlike knowledge injection methods that incorporate triplet-based structured knowledge [13], we work purely with free-text verbalizations generated from multimodal data. The retriever retrieves relevant perspective-conditioned passages, and the generator uses them to compose contextually rich, stakeholder-specific responses. This results in a system—PeRAG (Perspective-aware RAG)—that enables context-sensitive generation not just based on topical relevance but on the interpretive stance encoded in the input text passages. To the best of our knowledge, PeRAG represents the first instantiation of RAG tailored for multi-perspective decision support in urban governance contexts.

Although LLMs such as ChatGPT and GPT-3 [14] have shown great success in general-purpose generation tasks, their application in decision-making processes has been limited by a lack of specificity and contextual adaptation [15]. Generic outputs are often insufficient in high-stakes domains like urban planning, where conflicting group needs (e.g., between commuters, the elderly, and environmentally conscious citizens) must be mediated through nuanced communication strategies. Efforts like BLOOM [16] have underlined the importance of transparent, representative training data, particularly for multilingual settings. However, our implementation is currently focused on English language generation, which remains dominant in LLM infrastructure and evaluation. By operating entirely in English while incorporating multi-perspective reasoning, our approach can generalize to multilingual contexts in future iterations but already demonstrates strong utility in data-rich governance scenarios [17].

3. Methodology

Our methodology introduces a novel pipeline that bridges heterogeneous urban data and perspective-aware natural language generation using a tailored Retrieval-Augmented Generation (RAG) architecture. The following subsections detail our approach to homogenizing structured inputs, dataset preparation, verbalization strategies, system design, and evaluation.

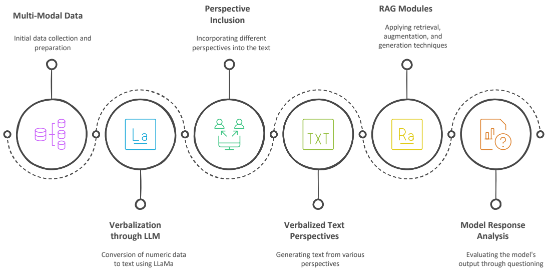
Multi-Modal Data
Initial data collection and preparation
Perspective Inclusion
Incorporating different perspectives into the text
RAG Modules
Applying retrieval, augmentation, and generation techniques
La
TXT
Ra
Verbalization through LLM
Conversion of numeric data to text using LLaMa
Verbalized Text Perspectives
Generating text from various perspectives
Model Response Analysis
Evaluating the model's output through questioning

Figure 1: PeRAG: Perspective inclusive pipeline with RAG

3.1. Homogenizing Heterogeneous Urban Data for RAG

Unlike conventional RAG systems that are designed to interface with a variety of knowledge representations—including tables, RDF triples, JSON schemas, and unstructured documents—our approach standardizes heterogeneous urban data into a unified format of unstructured textual narratives. This design choice fundamentally simplifies the retrieval mechanism and maximizes compatibility with LLM-based generation models. Rather than adapting the retriever to handle multiple data representations, we adopt a single retriever pipeline enabled by transforming structured data, including tables, geospatial indicators, and statistical measures, into natural language paragraphs. The resulting textual narratives are semantically enriched and explicitly crafted to reflect distinct analytical perspectives, ensuring that core domain-specific patterns are preserved while adapting the framing to match varied stakeholder viewpoints.

The homogenization approach offers several key advantages for urban policy applications. First, retrieval simplification is achieved through a unified representation that allows for a single dense retriever without requiring modality-specific modules, reducing system complexity and computational overhead. Second, our approach enables cross-modal comparability by facilitating reasoning across different data types, such as comparing demographics with transportation patterns through uniform verbal representations. Third, LLM compatibility is naturally reinforced by using natural language as both input and output, aligning with the intrinsic design of generative models and enabling seamless integration into query-response pipelines. Figure 1 outlines how PeRAG's components, multi-modal data, verbalization, perspective inclusion, RAG modules, and evaluation, integrate within the pipeline.

3.2. Dataset Description

The dataset comprises 7,019 urban data records covering Turin's geography, demography, and transportation systems from 2012 to 2019, offering a comprehensive longitudinal view of urban dynamics.

The data encompasses 3,850 census areas, which are portions of municipal territory organised in polygons, used by ISTAT[1] to divide the city into manageable, statistically meaningful areas. Demographic information about each census area is collected with respect to size and population distribution. Special attention is given to urban vulnerabilities, housing conditions, migration flows, and demographic changes in specific neighborhoods. Census areas can vary significantly in both size and demographic characteristics—they can be as small as a single street or encompass an entire residential block. For this reason, the census areas differ greatly from one another.

The census area is the smallest territorial unit used for analysis and is organized into 93 statistical zones. Statistical zones are aggregations of multiple census tracts and represent one of the intra-municipal territorial units into which the territory of the City of Turin is divided. In turn, the statistical zones are grouped into 9 districts - territorial subdivisions over which the local civil authority exercises its functions. This hierarchy of spatial units provides multiple levels of geographical granularity for analysis, enabling both fine-grained local insights and broader district-level policy evaluation. Additionally, the data for each census area is available for two reference years: 2012 and 2019, allowing for temporal comparisons across various dimensions. The dataset includes 31 structured features for each census-year tuple, systematically categorized into four primary domains.

Demographic information includes population density, gender distribution, age brackets, foreign residents, and the number of families, providing a comprehensive population profile. Additionally, the density of each demographic is calculated within a 500-meter buffer from the centroid of each census area. This approach accounts for the spatial distribution of density and makes the areas more comparable in terms of population concentration and access to services.

Public transport metrics include stop and line density, as well as connectivity indicators that measure how well each census area is linked to others in terms of accessibility and network coverage. Geographical identifiers encompass census codes, dimensions, statistical zones, district names, and boundaries that enable spatial analysis and policy targeting. Traffic and safety data document the number of accidents, vehicle involvement patterns, and the number of public transport incidents, supporting risk assessment and safety planning initiatives. This collection represents a significant expansion, enabling richer temporal and spatial analyses that capture urban evolution patterns and long-term policy impacts. The longitudinal scope allows for trend identification, seasonal pattern analysis, and evaluation of policy interventions over time.

The dataset was constructed by integrating multiple sources: all demographic data was obtained from the GeoPiemonte[2] portal, while public transport, traffic, and safety data were provided by Gruppo Torinese Trasporti (GTT)[3], which manages public transport services including urban, suburban, and extra urban routes, as well as tram and metro lines.

3.3. Perspective-Aware Verbalization of Urban Data

To enable retrieval over rich, interpretable textual data, we developed an Urban Data Verbalization System that translates structured urban records into fluent natural language narratives using large language models (LLMs). This system addresses the fundamental challenge of transforming quantitative urban data into qualitative insights that align with different stakeholder perspectives and analytical frameworks.

3.3.1. Verbalization

Our verbalization pipeline employs *LLaMA-3.1-8B* as the default model for processing numerical data and *LLaMA-3.2-11B-Vision* for processing heatmaps. The selection of these models allows us to maintain compatibility with other LLMs, ensuring both flexibility and reproducibility. We implement two primary verbalization strategies to balance generation quality with computational efficiency. Zero-shot verbalization allows the model to generate descriptions without specific examples, providing maximum creative freedom but potentially sacrificing consistency. Few-shot verbalization employs carefully curated single-shot examples that guide narrative style while preserving creative expression, resulting in more consistent and domain-appropriate outputs.

The system utilizes handcrafted prompts specifically designed to elicit structured yet non-hallucinatory summaries for each data record, ensuring factual accuracy while maintaining linguistic diversity. Two distinct prompt templates are employed: one for processing numerical tabular data using LLaMA-3.1-8B (see Table 6), and another for processing heatmap visualizations using LLaMA-3.2-11B-Vision (see Table 5). Complete prompt examples for both verbalization modalities are provided in Appendix C to ensure reproducibility. In both LLaMA configurations, generation control is achieved through carefully tuned parameters, including temperature set to 0.6 for optimal creativity balance, top-5 sampling at 0.9 for response diversity, repetition penalty of 1.2 to ensure

---

[1] National Institute of Statistics: https://www.istat.it/

[2] https://geoportale.igr.piemonte.it/cms/

[3] https://www.gtt.to.it/cms/

coherence, and the maximum token length is set to 512 for the 8B version and 1024 for the 11B-Vision version to support concise yet informative descriptions.

Each structured record is transformed into multiple narrative versions conditioned on distinct stakeholder perspectives. These include accessibility-oriented planning focusing on mobility and inclusion, safety and equity perspectives highlighting transportation risks and distribution fairness, and demographic inclusion addressing the needs of diverse populations. This multi-perspective approach ensures that verbalizations transcend generic summaries and address the specific analytical needs of different urban stakeholders. Table 3, presented in Appendix A, provides an example of this type of verbalization, illustrating both a general narrative and its corresponding multi-perspective version.

3.3.2. Quality Assessment and Validation

Unlike conventional LLM-generated general texts, which often suffer from loss of specificity, repetitiveness, or context ignorance, our perspective-aware narratives emphasize trends, deficiencies, and socio-geographic factors of particular interest to diverse urban stakeholders. The annotation protocol involved a systematic evaluation across four key dimensions: (1) contextual relevance whether the verbalization appropriately captures the urban context and stakeholder perspective, (2) information accuracy alignment between the verbalized content and source data, (3) coverage of information aspects completeness of perspective-specific elements in the verbalization, and (4) data factuality dealing with absence of hallucinations or fabricated information. Three expert annotators, including two postdoctoral researchers and one NLP researcher, independently evaluated a random sample of generated narratives for each dimension. Given the exploratory nature of this novel task and time constraints, a focused evaluation was conducted on a carefully selected subset of examples, with annotation disputes resolved through collaborative discussion among the research team. Their comprehensive assessment confirmed the validity, relevance, and framing alignment of perspective-aware verbalizations, providing empirical support for their use in downstream RAG generation tasks.

To mitigate potential ambiguities introduced during the natural language verbalization process, our approach incorporates several safeguards. First, the verbalization prompts explicitly instruct models to use exact numerical values without modification or approximation, preventing quantitative distortions. Second, the prompts restrict models from drawing conclusions, making assumptions, or interpreting data significance, thereby reducing interpretive ambiguity. Third, during the annotation process, evaluators specifically assessed verbalizations for information accuracy and data factuality, identifying instances where ambiguous phrasing might misrepresent the underlying data. Additionally, the multi-perspective approach inherently reduces ambiguity by providing explicit analytical framing, rather than generating generic descriptions that could be interpreted in multiple ways.

3.4. Perspective-Aware RAG (PeRAG)

PeRAG extends the traditional RAG paradigm to handle structured urban data through its verbalized form, creating a novel architecture specifically designed for perspective-aware policy support. The system integrates retrieval and generation components that work synergistically to provide contextually relevant and factually grounded responses to complex urban planning queries.

3.4.1. Retrieval Module

The retrieval module employs the *all-mpnet-base-v2* sentence transformer for dense vector encoding, chosen for its superior performance on semantic similarity tasks and computational efficiency. Text chunking is implemented using a token-based approach with a chunk size set to 500 tokens and an overlap of 50 tokens to ensure semantic continuity across chunk boundaries. This strategy ensures that semantically related content remains within the same retrievable segment, preserving coherence and relevance across retrieval operations.

The retrieval mechanism operates through cosine similarity-based semantic ranking with configurable top-k retrieval, defaulting to 5 results to balance comprehensiveness with computational efficiency. The system maintains comprehensive provenance metadata for complete traceability, enabling users and analysts to verify the source of retrieved information and ensuring accountability in policy-relevant applications.

3.4.2. Generation Module

The generation module utilizes *Gemma-3-4B-IT* as the default model while supporting any causal decoder-based large language model to ensure adaptability across different computational environments. The module processes user queries alongside retrieved perspective-aligned narratives using carefully engineered prompts that structure the input format as query plus perspective narratives.

Generation parameters are optimized for policy applications, with a temperature of 0.7 balancing creativity and factuality, and a 512-token limit ensuring brevity without sacrificing informational depth. The system demonstrates robust capability in responding to complex urban planning questions, supporting district-wise comparisons, demographic-transport correlations, safety and infrastructure assessments, and trend identification over temporal dimensions.

3.5. Implementation and System Efficiency

The full system is implemented in Python, leveraging PyTorch and Hugging Face Transformers for deep learning and natural language processing tasks, alongside SentenceTransformers for semantic retrieval capabilities. The implementation includes comprehensive batch processing capabilities with integrated performance monitoring to ensure scalable operation across large datasets. GPU acceleration with automatic device detection optimizes computational efficiency while maintaining compatibility across different hardware configurations.

The system architecture incorporates detailed logging for each transformation step, enabling comprehensive debugging and performance analysis. Key operational features include support for batch verbalization, which processes multiple records simultaneously; real-time querying capabilities for interactive policy analysis; and modular model swapping, allowing for easy adaptation to different language models or domain-specific requirements. This implementation approach ensures both research reproducibility and practical deployment feasibility for real-world urban policy applications. The source code for our PeRAG system, along with the various verbalization configurations, is publicly available at the following link[4]

4. Experimentation

Our experimental evaluation is designed to assess the effectiveness of perspective-aware verbalization and the overall performance of the PeRAG system in supporting urban policy decision-making. We conduct experiments across two primary dimensions: verbalization quality assessment and end-to-end system performance evaluation. All experiments are performed on locally deployed models to ensure data privacy and reproducibility, using NVIDIA GPUs for computational acceleration.

The experimental framework evaluates our system against several key research questions established in the introduction: the effectiveness of perspective-aware verbalization compared to general approaches, the comparative performance of zero-shot versus few-shot verbalization strategies, the utility of PeRAG for urban policy question answering, and the factuality and relevance of system outputs compared to general-purpose large language models.

4.1. Verbalization Evaluation Protocol

We conduct a systematic comparison between general verbalization, i.e., template-based approach, and perspective-aware verbalization approaches using our Turin dataset. General verbalization employs standard data-to-text generation without specific perspective conditioning, while perspective-aware verbalization generates targeted descriptions aligned with specific stakeholder viewpoints, including demographics-focused, transportation infrastructure-focused, temporal analysis, and deficiency assessment perspectives.

A random sample of 200 data records is selected for detailed verbalization analysis, ensuring representation across different districts, time periods, and demographic profiles. Our multi-modal dataset is processed through both zero-shot and few-shot verbalization strategies for each perspective type, generating a comprehensive corpus of verbalized descriptions for comparative evaluation.

For the verbalization quality assessment, two authors jointly annotated three representative examples in a structured meeting format, with any disagreements resolved through immediate discussion. While the limited sample size ($n = 3$) precluded formal inter-annotator agreement (IAA) calculation using Cohen or Fleiss' Kappa, the collaborative annotation process ensured consistency in evaluation criteria application. Future work will expand the annotation sample size to enable robust inter-annotator reliability metrics.

4.2. System Performance Evaluation

We develop a comprehensive set of 25 urban policy-oriented questions that span different complexity levels and analytical requirements. The question set includes factual queries about specific demographic or transportation metrics, comparative questions requiring cross-district or temporal analysis, analytical questions demanding trend identification and causal reasoning, and policy-oriented questions seeking recommendations based on data insights.

Questions are categorized by type (factual, comparative, analytical, policy-oriented), complexity level (simple, moderate, complex), and required perspective alignment (demographics, infrastructure, temporal, deficiency-focused). This categorization enables a systematic assessment of system performance across different query types and complexity levels.

System performance is evaluated against multiple baseline approaches to assess the contribution of our perspective-aware framework. These baselines involve querying general-purpose LLMs without access to urban-specific data. For this purpose, we use the *Gemini 2.0 Flash* and *GPT-4o Mini* models. Additionally, we evaluate RAG systems using general (non-perspective-aware) verbalizations under both zero-shot and few-shot configurations. Each baseline is tested using the same set of questions and evaluation criteria to ensure a fair and consistent comparison.

---

[4] Code and dataset are available at https://github.com/MasterHoracio/CLiC-it-HARMONIA.git.

4.3. Evaluation Metrics

In order to evaluate the performance of our proposed perspective-aware framework, as well as all the baseline approaches, we employ the Retrieval Augmented Generation Assessment (RAGAS) framework, specifically designed for reference-free evaluation of RAG pipelines [18]. This framework defines three main metrics. The first, *faithfulness*, measures whether the answer accurately reflects information that can be directly inferred from the given context. The second, *answer relevance*, evaluates whether the answer directly and appropriately responds to the given question, without being incomplete or redundant. Finally, the third metric, *context relevance*, assesses how well the context includes only the necessary information to answer the question, avoiding redundancy. For a detailed explanation, we refer the reader to the following paper [18].

5. Results

Table 1 presents the evaluation results for the different configurations considered. The first section of the table (rows 2 and 3) shows the results obtained by directly querying the LLMs without providing any additional context. It is important to note that the *faithfulness* and *context relevance* metrics could not be computed in this case, as both require access to the retrieved context. Nevertheless, the *answer relevance* scores reveal low performance for both models. This can be attributed to the fact that most of the responses were of the type *"I cannot answer the question due to lack of necessary data"*. Specifically, GPT-4o responded this way in 21 out of 25 cases, while Gemini 2.0 did so in 18 out of 25. Overall, Gemini demonstrated marginally better performance in this setting.

Additionally, Table 1 also compares the performance of general verbalizations using zero-shot and few-shot configurations. These results are shown in the second section of the table (rows 4 and 5). As can be observed, the *answer relevance* scores are higher than those obtained by the previously evaluated LLMs, which can be attributed to the incorporation of relevant information retrieved by the retrieval module. When comparing the general verbalization settings, we observe that the few-shot configuration outperforms the zero-shot setting across all three evaluation metrics, with an average improvement of 6%. This gain is likely due to the higher quality and greater level of detail present in the verbalizations generated under the few-shot configuration.

Finally, we present the evaluation results of our proposed PeRAG system. As shown, it achieves the highest scores across all three evaluation metrics, with an average improvement of 20% compared to the best-performing general verbalization configuration. Overall, the highest metric score was obtained in *faithfulness*, indicating that the responses generated by the PeRAG system effectively leverage information inferred from the provided context. On the other hand, the lowest score—both for PeRAG and previous configurations—was observed in the *context relevance* metric. This may be attributed to the diversity of information retrieved by the retriever module, which stems from the chunk partitioning strategy used. In particular, this strategy incorporated independent general and multi-perspective verbalizations for each district, zone, or census area.

Table 1
Evaluation results for the considered reference-free metrics. Reported values correspond to the average over the 25 evaluation questions. The prefixes ZS and FS indicate the *zero-shot* and *few-shot* configurations of the general verbalization.

Approach	Faithfulness	Answer R.	Context R.
GPT-4o	-	0.134	-
Gemini 2.0	-	0.163	-
ZS-RAG	0.685	0.582	0.166
FS-RAG	0.725	0.595	0.184
PeRAG	**0.793**	**0.626**	**0.272**

6. Analysis

To gain deeper insight into the performance of our proposed PeRAG pipeline, this section presents a quantitative and qualitative analysis of the generated responses. In particular, we conduct a comparative evaluation of the answers produced by the RAG system using the different types of verbalizations. For this analysis, we randomly sample three questions from our set of 25, focusing on the *demographic* and *transportation* perspectives. The selection of three questions for detailed BERTScore analysis was determined by several practical constraints. First, generating reference factual answers for comparative evaluation requires extensive manual verification against the original Turin dataset, which is a time-intensive process involving careful cross-referencing of multiple data sources and temporal dimensions. Second, as this represents an initial exploration of a novel task combining multi-modal verbalization with perspective-aware RAG, we prioritized depth over breadth in the qualitative analysis to thoroughly examine the mechanisms underlying performance differences between general and perspective-aware verbalizations. Third, the computational overhead of generating responses across all verbalization configurations and computing detailed semantic similarity metrics scales considerably with the number of questions analyzed. The three selected questions were chosen to represent different complexity levels and analytical requirements.

For each of these questions, we generate a reference factual answer by manually extracting and synthesizing the relevant information directly from the original Turin dataset. The reference answer generation process involves several systematic steps: (1) identifying the specific data fields and temporal dimensions required to answer each question, (2) querying the structured dataset to retrieve exact numerical values for the relevant census areas, statistical zones, or districts, (3) performing necessary aggregations or comparisons across the 2012-2019 timeframe where temporal analysis is required, and (4) formulating a concise factual response that accurately reflects the quantitative findings without interpretive bias. For instance, for questions involving demographic trends, reference answers include precise population counts, percentage changes, and specific demographic categories affected, all derived directly from the census data. This manual reference generation process, while labor-intensive, provides ground-truth answers that serve as reliable baselines for evaluating the factual accuracy and completeness of system-generated responses through semantic similarity metrics. We use the BERTScore metric [19], a widely adopted measure of semantic similarity between a generated text and a reference [20]. Finally, we present a discussion highlighting the strengths and weaknesses of the PeRAG pipeline compared to general verbalizations.

Table 2 presents the BERTScore evaluation results for the three randomly selected questions. The first section of the table (rows 2 and 3) reports the results for the general verbalizations, where the *few-shot* configuration achieves the highest scores across all BERTScore metrics. These outcomes are consistent with the trends observed in the reference-free evaluation metrics. The second section of the table shows the results for our PeRAG pipeline, which consistently achieves the best performance across all three metrics, further reinforcing the findings obtained through the reference-free evaluation.

We acknowledge that the BERTScore analysis based on three questions represents a preliminary assessment of semantic similarity performance, and the limited sample size constrains the statistical generalizability of these findings. The selection was necessitated by the substantial manual effort required for reference answer generation and verification against the multi-dimensional Turin dataset. Each reference answer requires careful extraction and synthesis of information across multiple data fields, temporal dimensions, and geographical units, followed by independent verification by domain experts. While these three questions provide initial evidence of PeRAG's superior semantic alignment with ground truth data, we recognize that broader systematic analysis is essential for robust conclusions. Future work will implement automated reference generation procedures and expand the evaluation to cover the complete 25-question set, enabling more comprehensive statistical analysis of semantic similarity performance across different question types, complexity levels, and analytical perspectives. Additionally, we plan to incorporate multiple semantic similarity metrics beyond BERTScore to provide a more comprehensive assessment of response quality and factual alignment.

Table 2

Evaluation results based on BERTScore. The columns report the macro-average recall, precision, and F_1 score across the three randomly selected questions. The prefixes ZS and FS indicate the *zero-shot* and *few-shot* configurations of the general verbalization.

Approach	Recall	Precision	F_1
ZS-RAG	0.818	0.831	0.821
FS-RAG	0.837	0.852	0.846
PeRAG	**0.851**	**0.873**	**0.862**

An important consideration in our verbalization approach is the management of potential linguistic ambiguities that could impact downstream RAG performance. Our analysis of generated verbalizations reveals that perspective-aware conditioning significantly reduces interpretive ambiguity compared to general verbalization approaches. For instance, when describing transportation infrastructure, general verbalizations might use ambiguous terms like 'adequate coverage' or 'reasonable accessibility', whereas perspective-aware verbalizations provide specific contextual framing, such as 'limited accessibility for elderly residents due to sparse stop density in residential areas'. This specificity not only reduces ambiguity but also enhances retrieval precision, as queries can be matched more accurately to relevant perspective-conditioned content. However, we acknowledge that some residual ambiguity remains inherent to natural language representation, particularly in cases where numerical thresholds are verbalized using qualitative descriptors (e.g., 'high density' vs. specific population counts). Future work will explore hybrid approaches that preserve exact numerical values alongside natural language descriptions to further minimize interpretive ambiguity.

To compare the outputs generated by our different configurations, Table 4 (included in Appendix B) presents a comparison between the response produced by our PeRAG pipeline and the one generated using the *few-shot* configuration of the general verbalization. This configuration was selected due to its strong performance in both the reference-free metrics and the BERTScore. Additionally, both responses are contrasted with a reference answer constructed from factual information. The question used in this analysis was selected from the set of three randomly chosen questions.

As shown in Table 4, the selected question involves a temporal comparison of demographic characteristics from 2012 to 2019. According to the reference answer, a population decrease is observed across most demographic groups, including males, females, minors, foreigners, and working-age citizens. In contrast, the only group that experienced population growth during this period was senior citizens.

When comparing these findings to the output generated by the PeRAG pipeline, we observe that it successfully identified the overall downward trend across multiple demographic groups, highlighting that the reduction was not evenly distributed. This aligns with the factual data presented in the reference answer. Moreover, PeRAG accurately captured the groups that experienced decline—such as the working-age population, minors, and foreigners—and correctly identified an increase in the senior population, consistent with the reference.

However, the PeRAG response emphasized the working-age population as the most affected category, whereas the reference answer pointed to foreigners. This discrepancy may be attributed to the nature of the multi-perspective verbalizations, which were generated at the level of census areas, statistical zones, and districts. Consequently, when retrieving information using the retriever module (configured with $k = 5$), it may not have captured a fully comprehensive view across all nine districts. This limitation has been corroborated by analyzing the retrieved chunks, where recalculating the values based on the retrieved verbalizations indeed showed that the working-age group experienced the largest decline.

Finally, Table 4 also includes the output of the general verbalization under the *few-shot* configuration. As shown, the response generated by the RAG system fails to clearly identify the downward trends across the different demographic groups as well as the upward trend for seniors. These results are consistent with those observed in the reference-free evaluation metrics. Moreover, although the response is factually correct, it does not address the perspective implied by the question, highlighting the importance of incorporating perspective-aware verbalizations. Similar to the PeRAG pipeline, the retrieved chunks in this configuration also exhibit limitations, indicating a potential area for improvement in future work.

7. Conclusion

This research demonstrates that multimodal urban data can be effectively verbalized through perspective-aware approaches to support policy-level interpretation, with our framework successfully processing over 7,000 examples across multiple analytical perspectives (RQ1). The comparative analysis reveals that few-shot verbalization strategies provide superior generation fidelity and perspective alignment compared to zero-shot approaches, despite increased computational overhead (RQ2). PeRAG, our lightweight locally-deployable RAG pipeline, effectively answers urban policy questions by leveraging these multimodal verbalizations as retrievable memory, ensuring data privacy while maintaining system responsiveness (RQ3). Human evaluation confirms that PeRAG exhibits superior factuality and utility compared to general-purpose LLMs in high-stakes policy scenarios, with domain-specific grounding providing enhanced accuracy and contextual relevance (RQ4). The framework establishes a reproducible methodology for transforming complex urban datasets into actionable policy insights, demonstrating that specialized, domain-grounded AI systems outperform general-purpose alternatives in critical decision-making contexts.

Limitations The various perspectives explored in this research, such as demographic, population, transportation, gender, and age, were derived from the dataset used in our evaluation. However, these perspectives do not incorporate public opinion. As ongoing work, we are expanding these perspectives through a research survey aimed at integrating viewpoints that reflect public opinion of citizens and stakeholders of Turin. The annotation protocol, while systematic, was applied to a limited sample size due to the exploratory nature of this novel task. The collaborative annotation approach, though ensuring consistency, does not provide quantitative measures of IAA. Future iterations of this work will implement larger-scale annotation studies with multiple independent annotators and IAA metrics to strengthen the evaluation framework. Additionally, we are working at enriching the evaluation framework. We plan to complement the reference-free evaluation metrics applied [21] by incorporating task-based evaluation protocols and comprehensive human evaluation strategies to better assess the practical utility of perspective-aware verbalizations in real-world urban planning contexts.

Acknowledgments

The research is conducted at the Department of Computer Science, University of Turin, Italy, and is funded by the “HARMONIA” project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 funded under the NextGenerationEU programme (PI: Viviana Patti).

References

[1] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: Concepts, methodologies, and applications, ACM Trans. Intell. Syst. Technol. 5 (2014). URL: https://doi.org/10.1145/2629592. doi:10.1145/2629592.

[2] Z. Li, J. Yang, J. Zhao, P. Han, Z. Chai, Pimr: Parallel and integrated matching for raw data, Sensors 16 (2016). URL: https://www.mdpi.com/1424-8220/16/1/54. doi:10.3390/s16010054.

[3] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. Ullah Khan, The rise of "big data" on cloud computing: Review and open research issues, Information Systems 47 (2015) 98–115. URL: https://www.sciencedirect.com/science/article/pii/S0306437914001288. doi:https://doi.org/10.1016/j.is.2014.07.006.

[4] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 423–443. URL: https://doi.org/10.1109/TPAMI.2018.2798607. doi:10.1109/TPAMI.2018.2798607.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. URL: https://aclanthology.org/P19-1656/. doi:10.18653/v1/P19-1656.

[6] P. Zhou, K. Gopalakrishnan, B. Hedayatnia, S. Kim, J. Pujara, X. Ren, Y. Liu, D. Hakkani-Tur, Think before you speak: Explicitly generating implicit commonsense knowledge for response generation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1237–1252. URL: https://aclanthology.org/2022.acl-long.88/. doi:10.18653/v1/2022.acl-long.88.

[7] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, H. Hajishirzi, Text Generation from Knowledge Graphs with Graph Transformers, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2284–2293. URL: https://aclanthology.org/N19-1238/. doi:10.18653/v1/N19-1238.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 1–16.

[9] S. Frenda, G. Abercrombie, V. Basile, et al., Perspectivist approaches to natural language processing: a survey, Language Resources and Evaluation 59 (2025) 1719–1746. doi:10.1007/s10579-024-09766-4.

[10] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: B. Williams, Y. Chen, J. Neville (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, 2023, pp. 6860–6868. URL: https://doi.org/10.1609/aaai.v37i6.25840. doi:10.1609/AAAI.V37I6.25840.

[11] X. Wang, P. Sen, R. Li, E. Yilmaz, Adaptive retrieval-augmented generation for conversational systems, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 491–503. URL: https://aclanthology.org/2025.findings-naacl.30/. doi:10.18653/v1/2025.findings-naacl.30.

[12] W. Chen, H. Hu, X. Chen, P. Verga, W. Cohen, MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5558–5570. URL: https://aclanthology.org/2022.emnlp-main.375/. doi:10.18653/v1/2022.emnlp-main.375.

[13] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Engineering Applications of Artificial Intelligence 133 (2024) 108166. URL: https://www.sciencedirect.com/science/article/pii/S0952197624003245. doi:https://doi.org/10.1016/j.engappai.2024.108166.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah,

J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 1–25.

[15] W. Liu, X. Wang, M. Wu, T. Li, C. Lv, Z. Ling, Z. JianHao, C. Zhang, X. Zheng, X. Huang, Aligning large language models with human preferences through representation engineering, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10619–10638. URL: https://aclanthology.org/2024.acl-long.572/. doi:`10.18653/v1/2024.acl-long.572`.

[16] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, et al., BLOOM: A 176b-parameter open-access multilingual language model, CoRR abs/2211.05100 (2022). URL: https://doi.org/10.48550/arXiv.2211.05100. doi:`10.48550/ARXIV.2211.05100`.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, et al., On the opportunities and risks of foundation models, CoRR abs/2108.07258 (2021). URL: https://arxiv.org/abs/2108.07258. `arXiv:2108.07258`.

[18] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020, pp. 1–41. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[20] M. Hanna, O. Bojar, A fine-grained analysis of BERTScore, in: L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, C. Monz (Eds.), Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, 2021, pp. 507–517. URL: https://aclanthology.org/2021.wmt-1.59/.

[21] D. Deutsch, R. Dror, D. Roth, On the limitations of reference-free evaluations of generated text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10960–10977. URL: https://aclanthology.org/2022.emnlp-main.753/. doi:`10.18653/v1/2022.emnlp-main.753`.

Table 3
PeRAG Verbalization Example: Multi-Perspective Data Transformation

Input/Perspective	Verbalization Output
Numeric Input	`year=2018; sez_cens=2535; stat_zone=75; area=145065.42; pop=192; perc_pop=2.69566E-05; perc_f=0.546875; per_immigrants=0.21875; perc_immigrants_F=0.135416667; perc_minor=0.130208333; perc_senior=0.223958333; n_stops=5; n_lines_stopping=5; perc_stops=0.026041667; perc_stops_per_line_stopping=1; perc_length_stopping=0.000167865`
General Verbalization	In 2018, census section 2535 in statistical zone 75 covered an area of 145,065.42 units, with a population of 192. The population percentage was approximately 0.0027%, of which 54.7% were female. Immigrants comprised 21.9% of the population, with 13.5% being female immigrants. Minors accounted for 13.0%, seniors 22.4%. There were 5 stops and 5 lines stopping, with stops representing 2.6% coverage, 100% stop-to-line ratio, and 0.017% length coverage.
Demographics Perspective	In 2018, census section 2535 in statistical zone 75 had a population of 192 individuals. Females accounted for 54.7% of the population. Immigrants represented 21.9% of residents, with female immigrants making up 13.5%. Children under 18 years old comprised 13.5% of the population, while seniors accounted for 22.4%.
Population Perspective	The total population of census section 2535 was 192, representing approximately 0.0027% of the overall population. The area covered by this section was 145,065.42 units, indicating a low population density given the large area size.
Geographical Perspective	Census section 2535 covers a substantial area of 145,065.42 units within statistical zone 75. Despite its size, the population remains small at 192 individuals.
Transport Infrastructure	Within census section 2535, there are 5 stops served by 5 lines. Stops cover 2.6% of the area, with an equal number of stops per line. The total length covered by stops corresponds to 0.017% of the area.
Immigration Perspective	Immigrants make up 21.9% of the population in census section 2535, with female immigrants comprising 13.5%. This suggests a significant immigrant presence within the local population.
Age Group Perspective	Minors account for 13.5% of the population in census section 2535, while seniors make up 22.4%, indicating a population with a notable proportion of older adults.
Gender Distribution	Females represent 54.7% of the population in census section 2535, showing a slightly higher female population compared to males.

A. Verbalization Examples

B. Example of Output Response

C. Verbalization Prompt Examples

To ensure reproducibility and transparency of our verbalization approach, we provide the complete prompt templates used for both numerical data (see Table 6) and heatmap verbalization (see Table 5).

C.1. Heatmap Verbalization Prompt (LLaMA-3.2-11B-Vision)

The following prompt template is used for generating narratives from population density heatmaps and other visual data representations:

C.2. Numerical Data Verbalization Prompt (LLaMA-3.1-8B)

The following prompt template is used for generating narratives from structured numerical urban data:

Table 4
Example of an output response to a randomly selected question, illustrating the outputs produced by the different verbalization configurations.

Question	From 2012 to 2019 the population in Turin has decreased. Considering all the census, does this decrease affect all the demographics? Which category is more affected and which one is less affected?
Reference Answer	Between 2012 and 2019, the population of Turin declined by 4.32%, and this trend affected most demographic groups. Both male and female populations decreased at nearly the same rate (4.34% and 4.31% respectively), indicating no significant gender disparity. The number of minors saw a notable decline of 5.96%, while the foreign (legally registered) population experienced the most significant drop, decreasing by 6.48%, together with the number of working-age citizens (6.28%). In contrast, the senior population was the only group to increase, growing by 2.05%
PeRAG Output	Analyzing the demographic shifts across Turin's statistical zones between 2012 and 2019, it is evident that the overall population declined during this period. However, the decrease was not evenly distributed across all demographic categories. The working-age population (19–64) experienced the most significant contraction, suggesting a notable loss of economic contributors. This trend was consistently observed across various districts. In contrast, the number of seniors slightly increased, indicating a growing aging population. The number of minors also declined, though to a lesser extent than the working-age group. Meanwhile, the foreign-born population registered a visible decline as well, potentially due to migration dynamics. In summary, while most demographic segments were negatively affected, the working-age group was the most impacted, whereas the senior population was the least affected, showing a modest increase.
FS-RAG Output	Analyzing the provided census data reveals a complex picture regarding demographic shifts between 2012 and 2019 in Turin's districts. While a precise overall population decrease for the entire city isn't evident from this dataset, several districts demonstrate notable changes. District 5 (specifically areas like Basse di Stura and Le Vallette) shows a consistent population size—around 41 residents in 2019, mirroring 2012—with minimal fluctuation within the 500-meter radius. Conversely, District 1 exhibits a clear increase in population size between the two years. In 2012, Census Area 799 (Politecnico) held 52 residents, expanding to 1,478 within a 500-meter radius. Similarly, Census Area 3209 (Politecnico) grew from 212 residents to 2,784. Looking at specific categories, the *minor* population appears significantly impacted by this growth, increasing dramatically in District 1 (from 32 to 423) compared to a smaller change in District 5. Seniors also show increased numbers in District 1 (47 to 652) while remaining largely stable in District 5. The foreign national population experienced a relative increase in District 1 (2 to 104) compared to a slight increase in District 5 (13 to 20). Therefore, while certain districts remained static, the overall trend indicates that the *minor* population is experiencing the most pronounced demographic shift, particularly within District 1, driven by expansion of the area analyzed.

Table 5
Prompt template used for map-to-text narrative generation

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert urban data analyst. Your task is to generate clear and precise narratives based
 on census and transport data for the city of Turin.
<|start_header_id|>user<|end_header_id|>

Generate a comprehensive narrative that analyzes and compares the {field_description.lower()}
across the statistical zones of Turin, based on the provided comparison maps.
The image displays comparison data for the years 2012 and 2019.
<|image|>
Your narrative must:
- Be concise, informative, and clearly highlight key patterns and trends in the
 {field_description.lower()}, considering both temporal changes
 (between 2012 and 2019) and within-year variations, where relevant.
- Provide a Top-summary for each of the following:
    - The most common patterns observed across zones.
    - Zones with the highest increases in values from 2012 to 2019
    (i.e., where 2019 value > 2012 value).
    - Zones with the largest decreases in values from 2012 to 2019
    (i.e., where 2019 value < 2012 value).
- Use the exact numerical values provided for each statistical zone—do not round, estimate, or
omit any data.
- Refrain from interpreting, inferring causes, or comparing with any external datasets or years
outside of 2012 and 2019.
Below are the statistical zones with their respective values for the selected field in 2012 and
2019:
{values}
Generated Narrative: <|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Table 6
Prompt template used for numeric-to-text narrative generation

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert urban data analyst. Convert census and transport data into clear narratives.
<|start_header_id|>user<|end_header_id|>

Generate a comprehensive, single-paragraph narrative about an urban area based on the following
numeric data.
The narrative must:
- Be concise, informative, cover all key aspects of the urban landscape, and limit to a single
 paragraph.
- Include and reflect the exact values as given in the Numeric Facts, without modification or
 approximation.
- Focus solely on describing the attributes defined in Field Descriptions, matching each field
 with its corresponding value.
- Avoid drawing conclusions, making assumptions, or interpreting the significance of the data.
- Avoid comparing the data to other entries, past values, or the example provided.
Unique Identifier: {row_context}
Field Descriptions: {field_description}
```