Real-time semantic enrichment multi-agent approach for urban mobility data

Selene Melo Andrade
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação (DCC)
Belo Horizonte, Brazil
selenemelo@dcc.ufmg.br

Prof. Antônio Alfredo Ferreira Loureiro
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação (DCC)
Belo Horizonte, Brazil
loureiro@dcc.ufmg.br

***Abstract*—Large Language Models (LLMs) have become ubiquitous in day-to-day life. Recent advancements concerning the integration of LLMs, multi-agent architectures and embedded systems have allowed the deployment of complex, data-oriented applications [1]. Leveraging such systems enables urban mobility applications to automatically collect data and learn dynamic patterns of vehicular traffic. Many urban disruptions, such as accidents, flash floods, roadworks and sporadic events, are announced in heterogeneous textual, visual and audio sources before becoming detectable in the traffic flow itself [2], yet existing semantic enrichment methods focus mostly on historical data and rarely exploit foundation models or agentic architectures for continuous, real-time enrichment [3], [4]. This work presents a model-agnostic, multi-agent pipeline that collects semantic data from social media and real-time streaming traffic content from radio broadcasts, transcribes and extracts structured urban events, and aligns them to vehicular trajectories to produce enriched and explainable routes. The pipeline is organized as a Mixture-of-Experts (MoE) agentic system communicating through versioned Pydantic contracts and coordinated by an orchestrator agent, with a dedicated reasoning agent that performs semantic interpretation under the ReAct paradigm and a human-in-the-loop validation protocol [5]. We instantiate the system for the city of Belo Horizonte, Brazil, using open and citizen-generated data, and evaluate extraction quality, faithfulness and operational cost. Preliminary results over a 24-hour deployment show that the deterministic traffic backbone contributes with 8.402 fully schema-valid, georeferenced incidents, while the LLM-based agents produce structurally valid events that remain grounded in their source text (94.9%–100% faithfulness, i.e., at most 5.1% hallucination). Against a blind, human-annotated gold standard, the news and social-media agents reach high F_1 scores of around 0.93, whereas the radio agent is bottlenecked by ASR model errors that propagate into over-extraction. The reasoning layer sustains this behaviour at low cost, using a lightweight 8B model at an average of 2.617 tokens per call and a median end-to-end latency of $25.5\,\mathrm{s}$ under a free-tier quota. These findings indicate that the multi-agent design yields structurally valid, well-grounded and auditable events at low cost, while clearly delineating where reliability must still be improved by using a more powerful model before unsupervised operation.**

***Index Terms*—Semantic trajectory enrichment; Large Language Models, Multi-agent systems; Real-time urban mobility; Explainable AI; Intelligent cities.**

I. Introduction

The comprehension of vehicular mobility patterns have become a paramount concern for a wide range of urban stakeholders. Big technology companies, such as Google, prompts a heavy effort into the visualization, understanding and annotation of mobility events. This task is now achievable due to the synergy among large-scale foundation models, the advanced semantic reasoning of LLMs, autonomous multi-agent frameworks, and the ubiquitous deployment of embedded systems in connected vehicles. This new paradigm allows the development of complex, data-oriented applications capable of collecting and processing high-velocity stream data in real-time. Within the scope of Smart Cities, this recent scenario evolves into a new way of planning transportation infrastructure, allowing the integration of the semantic intelligence powered by LLMs into the urban sensors mesh. Currently, vehicular traffic monitoring relies on Floating Car Data (FCD), the continuous and anonymized tracking of geographic coordinates, speed and timestamps generated by moving devices [6]. However, while FCD captures spatial traffic patterns, it operates reactively and lacks human-centric context. At the same time, critical urban disruptions (e.g., accidents, flash floods, roadworks, scheduled events) are often reported by citizens via real-time streams such as radio broadcasts and social media posts. Dynamically identifying, collecting, and representing urban trajectories through citizen-generated data leverages the rapid adaptation of foundation models, the semantic reasoning capabilities of LLMs, and the specialized execution of multi-agent systems. Hence, this approach generates explainable insights into traffic dynamics from a human-centric perspective, enabling context-aware mobility forecasting

Although existing Mobility Foundation Models (MFMs) in the literature are trained over billions of GPS and FCD data and are highly capable of capturing statistical patterns of movement, they face several performance boundaries imposed by the nature of their training data [4]. Since these models are trained with spatio-temporal data, they rely heavily on spatial tokenization, reducing urban space into discrete geographical coordinates. Although this mathematical abstraction effectively captures structural movement patterns, it overlooks all complementary informations about the locational and the human intent and behaviour over it. For instance, while a classical model can statistically detect a sudden drop in velocity vector of a moving object, it cannot infer whether it is caused by a routine bottle-neck or by an atypical congestion

due to a dynamic event, such as an unannounced street market or a cultural festival. Therefore, without real-time semantic enrichment those models lack the contextual awareness required to anticipated and explain how these human-centric anomalies impacts traffic behaviour. Besides, classical mobility analytics struggles with counterfactual reasoning. This happens since the model is pre-trained with historical data, failing to simulate how trajectories would shift under new constraints [7]. As published in the recent literature, classical deep learning predictors suffers from environment confound and spatial overfitting, which limits their capability of generalizing dynamic urban interventions [8].

To bridge this gap between raw spatio-temporal signals and the use of historical and reactive user-generated data, we propose an approach for continuous, real-time semantic trajectory enrichment. Our motivation relies on the fact that many mobility data with updated traffic constraints are published on social media and broadcasted by news radio stations. We argue that coupling foundation model reasoning with a structured multi-agentic pipeline can yields anticipatory and explainable mobility recommendations that provides advantages over reactive and FCD data. Therefore, this work aims to design and evaluate a system whose main objectives and expected contributions are:

1) Develop a model-agnostic pipeline organized as a mix-of-experts agentic system that continuously ingests user-generated urban data from social media and real-time streaming traffic audio from radio broadcasts;

2) Implement an event extractor framework that leverages OpenAI Whisper audio transcription coupled with Name Entity Recognition (NER) and LLMs to transform unstructured streams into typed urban events;

3) Design a semantic framework capable of interpreting, mapping and therefore enriching these dynamic events onto vehicular trajectories;

4) Validate the architecture through an evaluation protocol using the city of Belo Horizonte, Brazil as an example, assessing extraction quality, recommendation utility, and explanation faithfulness through a human-in-the-loop validation strategy.

The remainder of this article is organized as follows. Section II presents the research questions that drive this investigation. Section III reviews the background and the related work on semantic trajectory enrichment as well as the limitations and gaps of current real-time urban proposed ontologies in the state of the art. Section IV details the proposed methodology concerning the multi-agent approach for real-time semantic enrichment of trajectories addressed in this research. Section V establishes the results obtained and the discussions and evaluation methods around them. Section VI outlines future directions and the open issues to be addressed. Finally, Section VII concludes the article.

II. Research Questions

This work investigates whether a multi-agent architecture together with foundation model reasoning can deliver continuous and explainable semantic enrichment of urban trajectories from real-time, citizen-generated data. Whereas existing enrichment methods rely mostly on historical trajectories and static geographic data, our study is guided by the following research questions:

RQ1 Can a model-agnostic, Mixture-of-Experts multi-agent pipeline transform heterogeneous, unstructured citizen-generated streams (social media, news feeds, and broadcast radio) into structurally valid and georeferenced urban events, and semantically attach them to vehicular trajectories in real time?

RQ2 Is such an agentic enrichment architecture operationally and economically viable under tight resource constraints? Can an lightweight foundation models, coupled with a ReAct reasoning layer and a human-in-the-loop validation approach, sustain low-cost, auditable, and explainable route recommendations rather than black-box ones?

RQ3 To what extent are the events produced by the LLM-based extraction agents grounded in their source text? How faithful and free of hallucination is the extraction, and how does extraction quality vary across the different citizen-generated sources?

III. Background

In this section, we provide the literature background, the related works and the advancements and gaps in the state of the art regarding the semantic enrichment of urban mobility data.

Semantic trajectory enrichment is a well-established research field that combines knowledges from Computer Science areas, including Data Science, Computer Networks and Artificial Intelligence, with Transportation Engineering. State of the art methodologies predominantly categorize data as historical and real-time. Historical approached typically rely on post-processed trajectory informations stored in structured datasets. Because these datasets are easily obtained in web repositories, they account for most part of trajectory enrichment frameworks in the current literature. On the other hand, real-time data poses a major challenge in the field, since streaming data is harder to collect, structure and store due to its high velocity and intrinsic heterogeneity.

Literature on offline mobility data typically achieves semantic enrichment by dividing raw trajectories into sequential episodes of movement. These episodes are than merged each with static knowledge databases, such as DBpedia, using Linked Open Data (LOD) ontologies [9]. Beyond that, other strategies explore mobility pattern mining through clustering algorithms, such as DBSCAN, to process billions of historical FCD records and identify shared Point of Interests (POIs). Once these POIs are identified, probabilistic algorithm such as Markov Chains are used to align and infer the semantic meaning of each POI [10]. For instance, if a recurrent cluster of stops is detected around a specific location at 8 p.m., the model infers a high transition probability toward a POI corresponding to a restaurant, mapping a human intent as

dinning behaviour [11]. The literature also proposes modern techniques using Deep Metric Learning and Hierarchical Graph representation. Rather than mapping raw coordinates, these methods construct a network of Regions of Interests (ROIs) and train a deep neural network to learn the latent semantic proximity between these locations. Therefore, the network discovers hidden human-centric relationships that are not explicitly written in raw coordinates [12].

While these offline techniques benefits from having a global view of vehicle's historical path bypassing computational and latency constraint, they don't provide support for responsive Intelligent Transportation Systems (ITS). Since urban anomalies occurs frequently, it is necessary to come up with a method that captures the rapid change in urban dynamic environment.

The search for real-time semantic enrichment techniques remain a cutting edge topic in the urban mobility field. Currently, literature provides very few approaches to cover this task. Authors usually propose systems that uses flow processing engines, like Apache or Kafka, coupled with knowledge bases, such as LOD [13]. As FCD spatio-temporal data arrives in real-time, the system instantly makes the map-matching and cross-references the coordinate with POI APIs. Ontologies usually try to optimize map-matching latency and cost for massive data flows.

There are also recurrent neural networks (ST-RNNs) and spatio-temporal Transformers architectures that receives sequential current coordinates and tries to predict, at the same time, both the next location and the semantic meaning of the POI nearby [14], [15]. The result is an embedding space where time and space are translated into vectors that can carry semantic and behavioural informations based on users' history.

The best-known and widely used real-time data collection is mobile crowd sensing. This approach utilizes users' smartphone sensors and GPS to infer the transportation mean and the immediate surrounding context [16].

However, real-time approaches face significantly challenges regarding the computational cost of performing map-matching techniques. This step is indispensable for tracking current positions of a user in the city and consists of aligning raw and noisy GPS coordinates to the correct position of the street on the map. This process avoids the snap-to-road problem, which refers to the inability to accurately determine which road a vehicle is actually on. It happens due to the GPS inaccuracies and can cause several position related problems such as ghost turns, trajectories disconnection, overpass confusion and odometer miscalculations [17]. Despite recent advancements in streaming architectures designed to mitigate asynchronous data ingestion, existing real-time enrichment methods still struggles with the dynamic semantic of the urban environment [18]. Most literature frameworks enrich streaming FCD by aligning raw coordinates with static geographic primitives or pre-existing POI databases [19]. However, the dynamic urban semantic still an open field of research. For instance, if a street is closed due to an unpredictable event, the traditional flow engine would not capture that context, since the event does not exists in static geospatial datasets. Regarding data types, structured and numerical data, such as flow matrixes, GPS logs and traffic counters are most commonly used. Unstructured data, like audio, images or posts on social media are seldom used. Yet, they carry important contextual meaning about their surrounding environment. Beyond that, when conventional geographic application, such as Google Maps, tracks a congested flow, they recalculate the path without providing human-centric explanations regarding why to deviate and why the proposed route is better than many other possible options. Those systems usually works as a black-box and lacks a contrafactual reasoning layer combined with natural language for human-centric, explainable routing.

IV. Methodology

This section details the architectural design and implementation of the proposed pipeline, as well as the underlying algorithmic decisions. The framework is conceived as a model-agnostic, multi-agent architecture leveraging a Mixture-of-Experts (MoE) approach, in which an orchestrator agent coordinates a set of specialized agents. Each specialist agent is responsible for the real-time collection, transcription, and interpretation of multimodal semantic data. The pipeline is implemented in Python and relies on publicly available web datasets and APIs, including OpenStreetMap (OSM), the TomTom API, and official mobility data from the city of Belo Horizonte (PBH), Brazil, where the system is instantiated. In summary, the methodology is structured into five core stages: **(1) Multi-source data collection; (2) Raw trajectories preprocessing; (3) MoE agentic collection, extraction, and structuring of heterogeneous semantic data; (4) Semantic aggregation and representation of the enriched trajectories;** and **(5) Evaluation and validation strategies comprising an reasoning agent, a human-in-the-loop protocol, and empirical metrics**. The final outputs consist of semantically enriched trajectory datasets provided in both tabular and human-readable natural language formats. A representational schema of the proposed methodology can be seen in the figure bellow.

A. Data collection

This step involves collecting the heterogeneous signal that feeds the system. The data gathered at this stage are contextual urban events that may affect mobility. Those informations are collected from four different types of sources. The first one is structured traffic data, obtained from TomTom's Traffic Incidents RESTful API over the bounding box of Belo Horizonte. Each incident provides its type, severity, local geometry, affected road stretch, and time stamps. At this stage, we also perform some basic preprocessing procedures to reduce data payload size and API quota consumption. From Tomtom, we request only the fields that are strictly required by the event contract (incident type, severity, local geometry, origin and destination road references, and timestamps) and discard duplicated incidents. Besides, instead of collapsing each incident to a single representative coordinate point, we preserve the full width of the affected road segment to retain the spatial extent

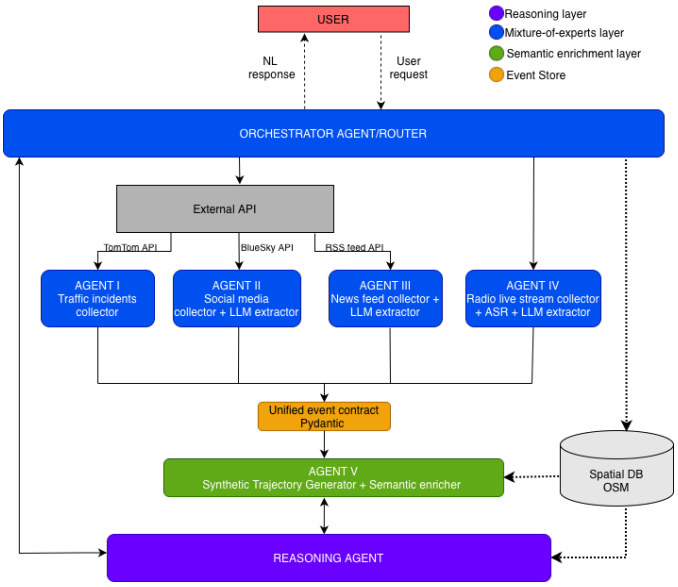

Fig. 1. Proposed multi-agent semantic enrichment pipeline.

of the disruption for a finer alignment during the enrichment phase. The second data source comprises citizen signals, e.g., posts retrieved from Bluesky social network. At this stage, we search for any post that mention the city of Belo Horizonte and then we apply filters to select only those mentioning traffic-related events to append to the unified event contract. The third data source are obtained from RSS feed (Rich Site Summary), which is a text format for real-time distribution of information over the Internet. The RSS technology allows internet users to subscribe in websites that provides RSS feeds and it is widely used by news companies such as CNN and BCC. Nowadays most of websites provides RSS feed and they serve many purposes including marketing, bug reports and any activity that involves updating or constantly publishing contents [20]. With this format of data retrieval, we collect news feed from different sources of traffic news in Belo Horizonte, such as BHTrans data, PBH data and other news feed retrieved from Google.

Finally, the fourth and last data source is broadcast traffic radio audio, continuously captured from a local news station filtered to collect only mobility-related and cultural events informations and later transcribed and structured by a specialist LLM agent.

B. Raw trajectories preprocessing

An important design decision concerns to the raw trajectory obtaining process. Raw GPS trajectories from OSM exhibts several issues such as very short duration, mostly historical trajectories, low sampling rates, or noisy data points making them not suitable for the real-time representation. Besides that, individual, real-time vehicular trajectories are neither publicly available nor free to use, as they constitutes personal data and are under privacy regulations or treated as a proprietary asset by navigation providers, requiring payment to get access to. Therefore, to avoid this issue, the trajectories the system enriches are generated synthetically. We sample origin-destination (OD) pairs and route them over the Open Source Routing Machine (OSRM) on an OpenStreetMap extract of Belo Horizonte obtaining, for each pair, the optimal segment together with its length and estimated duration. By distributing the duration along the segment, a timestamp is assigned to every vertex, yielding a trajectory in the spatio-temporal form $T = \langle (x_i, y_i, t_i) \rangle$ introduced in Section III.

To avoid choosing OD pairs at random, we anchor them on the collected incidents. Origin and destination are, therefore, chosen from the coordinates of TomTom events so that the resulting routes traverse the regions where context actually exists. This decision design is intended as a proof of concept and shows that a small, well-chosen set of trajectories is sufficient to validate the enrichment method, whereas evaluating it at scale, or over real traces, is left as future work since it depends on paid licences. It is worth acknowledging the limitations of this strategy. The OD pairs reflect plausible rather than observed travel demand, and the trajectories are simulated rather than collected. Publicly available real traces, such as the OpenStreetMap GPS trace repository, constitute a possible alternative but are a sparse and biased sample of urban mobility and are therefore unsuitable as a representative volume. These choices are revisited in the discussion of limitations.

C. Multi-agent (MoE) System

Specialist Agent 1 - Traffic Incidents Collector: Agent responsible for automatically polling a RESTful API that provides traffic informations over the bounding box of the city of Belo Horizonte, where the pipeline were instantiated. The API used was TomTom Traffic Incident due to its quick refresh time (around 20 seconds) and also because it is the only free API that provide such informations. To reduce payload and API quota consumption, the agent only request traffic-related events, such as incidents described by type and geometry, events with its description, magnitude of delay and date and time logs. Each extracted event are normalized to the unified event contract, the Pydantic schema. The output is a .JSONL file appended with the extracted events.

Specialist Agent 2 - Social Media Collector: Agent specialized in automatically collecting traffic-related social media posts that explicitly mention the city of Belo Horizonte. The API used for this task was BlueSky API chosen since it offers free programmatic access. The approach we used was to define a set of mobility keywords such as accidents, congestion, roadworks, events, combined with city markers. Each retrieved post goes through a filter that discards content unrelated to local mobility and the remaining posts are passed to a Large Language Model to extract the structured traffic events. The extracted location is geocoded and kept only if it falls within the bounding box of Belo Horizonte.

Specialist Agent 3 - News feed Collector:

This agent is responsible for collecting RSS feeds, a data format available on the Internet that automatically updates websites contents. In our case, we rely on Google News RSS, gathering incidents in Belo Horizonte from sources such as BHTrans, the PBH feed, Minas Gerais Globo News (G1), among others. Since it draws from official agendas and local press, this agent provides reliable sources, specially for scheduled events (fairs, marathons, matches, etc) which the radio and social-media agents tend to reports with more noise. As with the other unstructured sources, each item is filtered for relevance, processed by the LLM to extract structured events, and geocoded within the bounding box of Belo Horizonte before being stored.

Specialist Agent 4 - Live Radio Stream Collector:
Agent responsible for collecting and extracting events from unstructured audio streams. To do so, as a first step, an audio segmentation process consumes the live radio stream, captured by a web live transmission of the radio. HTML protocol provides the URL that directly point to the real time stream sequence. The audio is then segmented into 60 seconds chunks so the LLM agent can perform the transcription quickly, as the computational time to transcribe the audio grows with the number of words to transcribe. after processing step, each chunck is immediately removed from disk so that no raw audio is persisted, addressing the privacy concerns. In the second step, using Foundation Models, each completed chunck is transcribed by an Automatic Speech Recognition (ASR) model, OpenAi Whisper, constrained to Portuguese nd to deterministic decoding, producing a text rending of the traffic bulletin. In the third step, the text produced is therefore written by an LLM model, Groq LLama 3b, to extract the information regarding to traffic incidents. The incidents are then normalized to the pydantic schema to compose the dataset used in the enrichment step and appended in a .JSONL file.

Orchestrator Agent:
Agent responsible for coordinating the collector agents and supervising their execution. Acting as the manager of the multi-agent system, the orchestrator agent launches each collector agent as an independent, supervised process and manages its entire life cycle. Because every collector emits events under a common schema, the orchestrator treats them uniformly. Supervision is structured around three main goals. The first is the crash recovery, where the orchestrator automatically restart an agent if it terminates unexpectedly through an exponential backoff strategy and a bounded number of restarts within a time window. The second is the stall detection and it is triggered when an agent remains alive but stops producing output. This scenario is characterized as a silent failure and is detected by continuously monitoring the staleness of each agent's activity. The third one is the graceful shutdown and it happens upon a stop signal or reaching a configured run duration, so the orchestrator propagates the shutdown to every child agent in an orderly fashion. Beyond those main tasks, the orchestrator also manages auxiliary concerning the real-time long-term data collection, such as preventing the host machine from sleeping, persisting agent logs and tracking process identifiers for monitoring.

Reasoning Agent:
Agent responsible for the semantic interpretation of the entire system, transforming raw and unstructured natural language input into structured mobility events. It is the reasoning component to which the collector agents primarily sends their output, ensuring that all natural language understanding happens in one place and under one unified contract. For each input text, the agent prompts a LLM, for instance, through the Groq API, using Llama-8B for short texts and a larger model for noisier transcriptions., and reasons over the content to produce a structured event. The reasoning approach comprises **(i)** classifying the event into a predefined mobility-related taxonomy, such as accidents, roadwork, road closure, congestion, scheduled event, adverse weather, special operation, among others; **(ii)** identifying the mentioned location expressed in text and geo-localizing it under the city map; **(iii)** resolving temporal expression such as "now", "until the weekend", "until tomorrow", among others into absolute time windows, respecting the temporal reference as the moment the text was produced; **(iv)** estimating the event severity and **(v)** assigning a confidence scored in $[0, 1]$, based on whether the reported location is vague or ambiguous and if there is an explicit source from where the reported event was collected. The output is validated against a strict schema, for instance, using the Pydantic framework, which constitutes the formal contract consumed by the downstream Orchestrator and the enrichment stages. To remain robust during unattended operation, the agent transparently retries on provider rate limits and can fail gracefully on a single problematic item without interrupting the collection loop.

D. Semantic aggregation and representation of the enriched trajectories

The reasoner agent performs semantic aggregation using the ReAct pattern to generate both reasoning traces and task-specific actions in an interleaved manner [5]. This paradigm couples decision-making with textual justifications, naturally producing auditable traces and explanations for the end user. To do so, the agent turns the event stream and the generated trajectories into a single enriched artifact. Incoming events are accumulated in a continuously update store and deduplicated using the persistent identifies maintained by the collectors, so that repeated reports of the same incident are minimized, avoiding the inflation of the context. Each event is grounded to a geographic coordinate and to a temporal window, and is then attached to the trajectory through a spatio-temporal join. With this approach, an event enriches a trajectory segment where

their spatial proximity is below a distance threshold and their temporal window overlap. There is:

$$\mathrm{enrich}(s,e) \iff \mathrm{dist}(s,e) \le \delta \wedge [t^e_{\text{ini}},\, t^e_{\text{fim}}] \cap [t^s_{\text{ini}},\, t^s_{\text{fim}}] \neq \varnothing. \tag{1}$$

Where s denotes a vehicle trajectory and e denotes an urban event collected by one of the agents. The predicate $\mathrm{enrich}(s,e)$ is true if and only if the event e is considered relevant to the trajectory s, and in that case s is semantically enriched with e. The enrichment procedure requires the conjunction ($\wedge$) of two conditions. The first, $\mathrm{dist}(s,e) \le \delta$, is a *spatial* criteria where $\mathrm{dist}(s,e)$ is the smallest distance (computed with the haversine formula over geographic coordinates) between any point of the trajectory and a representative point of the event, and δ is a configurable proximity radius (e.g., $\delta = 120\,\mathrm{m}$). The event is spatially relevant only if it occurs close enough to the path actually travelled. The second condition is a *temporal* criteria expressed as the intersection of two time intervals: $[t^e_{\text{ini}}, t^e_{\text{fim}}]$ is the validity window of the event (its start and end instants, as resolved by the Reasoning Agent), while $[t^s_{\text{ini}}, t^s_{\text{fim}}]$ is the time span of the trajectory. The requirement $[t^e_{\text{ini}}, t^e_{\text{fim}}] \cap [t^s_{\text{ini}}, t^s_{\text{fim}}] \neq \varnothing$ states that these two intervals must overlap, which is, the event must be in effect while the vehicle is traversing the region. Only when an event is both spatially close and temporally concurrent with the trajectory the enrichment is performed.

The result materializes the Multiple Aspects Trajectory (MATs) model introduced in Section III, where the enriched trajectory is serialized as an extended GeoJSON document in which each segment carries the typed events that affect it, together with their sources and confidence scores. In parallel, in our approach, every event is encoded by an embedding model and indexed in a vector store, enabling the reasoning agent to retrieve semantically related events by similarity rather than by exact match. For inspection and manual annotation, the event stream is also exported to a tabular spreadsheet (.xlsx format) in which each event occupies one row and each field one column.

E. Evaluation and validation strategies and empirical metrics

Validation is built upon a reasoning agent integrated with a human-in-the-loop approach. Operating under the ReAct paradigm, the agent consumes the enriched trajectory, alternating reasoning steps with tool invocations, vector store semantic retrieval, and structured queries over the trajectory ontology. Ultimately, it reasons over the events on the trajectory accompanied by a structured justification. This justification details the evaluated events, their sources, the aggregated confidence, and the natural-language reasoning chain linking evidence to the final decision. Furthermore, the architecture supports both cloud-hosted and locally quantized deployment, enabling the empirical measurement of the trade-off between explanation quality and latency.

The evaluation pursues two complementary goals. The first assesses extraction accuracy by benchmarking the events automatically generated by the live radio collector against a human-annotated reference corpus of radio bulletins and official records, measuring precision, recall, and F1-score per event type. The second evaluates explanation quality through automated faithfulness and groundedness metrics, using an LLM-as-a-judge approach. Finally, the human-in-the-loop protocol closes the cycle by validating and correcting the events and explanations, human evaluators provide the supervisory feedback to calibrate confidence thresholds and refine prompts, iteratively improving system reliability.

Given our multi-agent approach, the empirical validation operates at two levels. First, each agent is evaluated separately to measure its specific performance. Second, the entire system is assessed as a whole to capture the end-to-end effectiveness of the architecture. Bellow we describe de metrics used.

*1) **Word Error Rate (WER)**:* Applied exclusively to the live radio collector agent (agent 4), WER measures the quality of the Automatic Speech Recognition (ASR) stage powered by the Whisper model. It compares the generated transcript against a manually produced ground-truth reference transcript. WER is defined as the ratio of word-level edits (insertions, deletions, and substitutions, computed via Levenshtein distance) to the total number of words in the reference:

$$\mathrm{WER} = \frac{S + D + I}{N} \tag{2}$$

Where S is substitutions, D is deletions, I is insertions, and N is the number of words in the reference. It reflects a structural importance over our approach, since the live radio agent entire extraction pipeline operates downstream from the transcribed text, any ASR error propagates into the extracted events. Reporting the WER explicitly defines the upper bound of quality achievable by this data source.

*2) **Pre-filtering Rate**:* This metric is also specific to Agent 4 (live radio collector), and measures the proportion of audio segments that, once transcribed, are classified as relevant to mobility and subsequently forwarded to the LLM. This metric is critical because a large portion of radio programming such as music, commercials and jingles contains no traffic events. The pre-filtering rate quantifies the volume of data discarded prior to LLM processing, demonstrating the API quota and computational cost savings provided by the filter.

*3) **Precision, Recall, and F1-Score**:* Applied to the three unstructured data agents (Radio, Social Media, and News), these metrics evaluate event extraction quality by comparing the Reasoning Agent's output against a blind and manually annotated ground truth, which was annotated by three human volunteers. The annotator records true events by reading only the source text, without access to the model's output. Precision indicates the proportion of extracted events that are true positives, Recall indicates the proportion of actual events captured by the model and the F1-score synthesizes the balance between both:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}} \tag{5}$$

Let TP be True Positives, FP be False Positives and FN be False Negatives. An extracted event is considered a True Positive (TP) value when its local matches, by textual overlap, the location of a goals standard event not yet consumed. The remaining cases define False Positives (FP) and False Negatives (FN). The alignment between an extracted event and a reference event is determined via location matching. These constitute the main reliability metrics of the system, directly quantifying the trustworthiness of the events feeding the enrichment pipeline.

*4) **Wilson Score Confidence Interval**:* Computed alongside each precision and recall value. Rather than reporting only a point estimate, the Wilson interval provides a plausible range for the proportion given the sample size. Its inclusion is essential for methodological transparency since the annotated ground-truth dataset is relatively small, confidence intervals communicate uncertainty to the reader and prevent the over interpretation of minor differences between sources.

*5) **Faithfulness (Source−Text Alignment)**:* Applied to the LLM-based agents (Radio, Social Media, and News), this metric measures the fraction of events whose text evidence, that we saved as *source_excerpt*, appears literally within the source text. The complement of this rate (1_faithfulness) estimates the model's hallucination rate. This is critical in LLM extraction pipelines due to the risk of generating fabricated information and verifying that each event is anchored in an actual source quotation enhances the traceability and auditability of the outputs.

*6) **Token Cost**:* Applied to the Reasoning Agent and broken down by the specialists agent (Radio, Social Media, and News), this metric aggregates the total number of input and output tokens consumed during LLM API calls, reporting averages per call and per extracted event. Agent 1 (Traffic Incident collector) does not incur this cost as it is deterministic. This metric documents the computational and financial operational costs required to evaluate pipeline scalability.

*7) **Latency and Throughput**:* Measured at the Reasoning Agent level, these metrics record the execution time per model call (mean, median, and variance) alongside the number of source texts processed per second. They are crucial for characterizing system viability under continuous, near-real-time operation, which is the specific deployment regime for which this pipeline was engineered.

V. Results and discussion

A. Per-agent results

1) Agent 1 — Traffic Incident Collector: Being a deterministic collector, this agent incurs no LLM cost. This agent only captures the already structured event available at TomTom Traffic Incident API. It collected $8{,}402$ incidents, all schema-valid and fully georeferenced via the geometry returned by the provider. Its severity labels are derived deterministically from the provider's delay magnitude, and its event types are dominated by congestion ($8{,}197$ of $8{,}402$), as expected from a continuous traffic-incident feed. This agent therefore acts as the high-volume, high-reliability structured backbone that complements the noisier unstructured sources. However, it captures vary few semantic informations regarding to scheduled events and roadworks informed in the official portals.

2) Agent 2 — Social Media Collector: The agent produced 11 events over the 24 hours of execution, all of which (100%) were schema-valid, fully faithful to the source text (100%), and successfully geocoded within the Belo Horizonte bounding box (100%). Its mean confidence was the lowest among all sources (0.50, median 0.30), consistent with the noisier and more colloquial nature of social media posts, where locations are frequently vague. However the number of collected events was small, since we used only one type of social media API (BlueSky API) since it is the only free-to-use API. This was a trade-off we had to deal with because, at the moment the pipeline was developed and tested, it was not possible to choose for paid social media APIs due to limited resources.

3) Agent 3 — News Feed Collector: The news agent extracted 107 events over the 24 hours of continuous execution with 100% schema validity, 97.3% faithfulness (2.7% hallucination), and 100% geocoding success. It also exhibited the highest mean confidence (0.84, median 0.90), in line with its role as the most reliable source for scheduled events, whose descriptions are more explicit and structured than spoken or social-media reports. The reliability of this sources is explained since they are official news feed websites, such as the BHTrans and PBH website, the local agenda of the city of Belo Horizonte.

4) Agent 4 — Live Radio Collector: Over the captured stream, with the code being activily collecting data for uninterrupted 24 hours, the agent transcribed $1{,}447$ audio chunks, of which only 119 (8.2%) passed the mobility pre-filter and were forwarded to the LLM. The remaining 91.8% (music, advertisements, jingles and other not mobility related content) were discarded before any model call. This pre-filtering rate is the single source of LLM-cost savings in the pipeline. From the processed segments, 98 events were extracted, 95.9% of which validate against the current structured schema. The 4 invalid cases correspond to an event type, `manutencao_metro`, added to the schema after this collection, since this type of incident was capture many times and was not predicted before the implementation as an event type. We can classify this category of error as a schema drift, where it was necessary to adapt the schema to capture more types of events. This result means that a high rate of the events that has passed through the mobility filter were in fact extracted, meaning that the filter performed well. Faithfulness reached 94.9%, implying an estimated hallucination rate of 5.1%. The mean self-reported confidence was 0.74 (median 0.80). The Word Error Rate of the ASR stage is reported as *pending*, as it requires a manually

transcribed reference corpus that is still under construction.

TABLE I
PER-AGENT EXTRACTION QUALITY AND OUTPUT CHARACTERIZATION.

Agent	Events	Schema	Faithf.	Confiability (mean/med.)
A4 Radio	98	95.9%	94.9%	0.74 / 0.80
A2 Social	11	100%	100%	0.50 / 0.30
A3 News	107	100%	97.3%	0.84 / 0.90
A1 Traffic	8402	100%	n/a	n/a

B. Reasoning Agent - Cost, Latency, and Throughput

Across the unstructured agents that delegate to the reasoning layer of the MoE approach, the Reasoning Agent served 285 LLM calls totalling 745,847 tokens, with an average of 2,617 tokens per call (Table II). The news agent dominates token consumption, both in volume of calls and in output length per call (439 completion tokens on average, against 181 for radio and 98 for social), reflecting longer and more event-dense source texts. It is worth noting, however, that the token cost might be overwhelmingly driven by the prompt (88.6%), whose schema and few-shot instructions are sent on every call regardless of the input length. The limited-resource constraint once again imposed a trade-off between model capability and token usage. Since paying for a more capable model was not viable, we had to choose one that balances this trade-off. As a consequence, all reasoning is performed by llama-3.1-8b-instant, a lightweight 8B model whose larger free-tier quota makes the dominant prompt cost sustainable, at the expense of the deeper reasoning a larger model would offer on the longer, event-dense news texts.

This choice also shaped latency and throughput. On the free tier, throughput was bounded not by the model but by the provider's rate limits: the 285 successful calls triggered 321 rate-limit back-off events (each pausing the pipeline for 20 s), concentrated on the news agent (217 events) owing to its higher call volume and longer prompts. Throughput was therefore governed by quota-induced waiting rather than by raw inference speed, so the effective end-to-end latency reflects scheduling around the rate limit far more than the per-call inference time.

TABLE II
TOKEN COST OF THE REASONING AGENT, BROKEN DOWN BY DELEGATING AGENT.

Source agent	Calls	Total tokens	Tokens/call
A4 Radio	112	266,451	2,379
A3 News	140	410,285	2,931
A2 Social	33	69,111	2,094
Total	**285**	**745,847**	**2,617**

The median end-to-end latency per call was 25.5 s (mean 23.5 s, $\sigma = 14.8$ s), at an average cost of approximately 1,900 tokens per extracted event. We stress that this latency was measured under the provider's free-tier quota and therefore includes rate-limiting back-off. This should be read as the operational latency under quota constraints rather than the raw model inference time. Removing the quota ceiling, or adopting the locally quantized deployment supported by the architecture, is expected to reduce this figure substantially and is left as future work.

C. End-to-end extraction quality

Table III reports precision, recall, and F1 against the blind human-annotated gold standard, with 95% Wilson confidence intervals. Because the saved output of the radio agent covers only a subset of the annotated corpus, we can report two observations. The over all annotated items, where unprocessed items are unavoidably counted as false negatives, and restricted to items the pipeline actually processed. The coverage of the saved corpus was 100% for social, 75.6% for news, and only 28.6% for radio, which explains the gap between the two views for the radio source.

TABLE III
EXTRACTION QUALITY VS. BLIND GOLD STANDARD

View	Source	Precision	Recall	F1
(a) All	News	78%	52%	62%
	Radio	27%	28%	27,4%
	Social	88%	100%	93%

D. Discussion

The results expose a contrast between the structured and unstructured sources. The deterministic traffic agent is, by construction, fully reliable in schema validation and georeferencing, providing a dense and trustworthy event backbone. Conversely, the LLM-based successfully mitigate the two main concerns with generative extraction. The first is the *structural validity* ($\geq 95.9\%$ schema-valid) and the second the *groundedness* (94.9%–100% faithfulness with at most 5.1% hallucination). In other words, when the model emits an event, it is almost invariably well-formed and anchored in a real source quotation, which contributes with the auditability of the pipeline itself.

Nevertheless, the end-to-end extraction quality against the blind, human-annotated gold standard reveals an overall F1 of approximately 0.30 on radio processed items. This bellow expectation performance highlights the main bottleneck of the pipeline. Two factors dominate this behaviour. First, precision acts as the primary constraint, indicating that the model over-extracts by producing events that human annotators did not consider genuine mobility incidents or with erroneous extractions. Second, we observed that utilizing a medium-sized model for ASR (Automatic Speech Recognition) transcription introduces significant textual and phonetic noise. The LLM subsequently attempts to extract events from these corrupted transcripts, leading to misinterpretations and false positives. Upgrading to a more precise ASR model is expected to minimize transcription errors, thereby reducing false positives and consequently driving higher precision, recall, and F1-scores.

These findings motivate future improvements that are already supported by the system architecture, that is, (i) integrating a more powerful ASR model to ensure high-fidelity tran-

scriptions and prevent downstream textual misinterpretations; (ii) tightening precision through prompt refinement and optimizing confidence thresholds calibrated via human-in-the-loop feedback and (iii) transitioning from free-tier deployment to an unconstrained cloud API or a locally quantized model to obtain better overall results that depends on models' capabilities. To summarize, the system demonstrates that the proposed multi-agent design produces structurally valid, well-grounded, and auditable events at low cost, while clearly delineating where reliability must be improved prior to unsupervised operation.

VI. Future directions

The results reported characterize the present system as a proof of concept and, at the same time, expose a clear routine for turning it into a deployable service.

Improving extraction reliability: The end-to-end evaluation identifies the precision metric as the dominant bottleneck, driven almost entirely by the radio agent, where Automatic Speech Recognition errors propagate downstream and induce over-extraction. Two directions follow naturally from the architecture's model-agnostic design. First, replacing the constrained transcription and extraction models with a more capable foundation model is expected to reduce transcription noise and the spurious events it generates, raising precision, recall and F_1 jointly. Second, because the reasoning agent already emits a calibrated, per-event confidence score, a confidence threshold tuned with additional human-in-the-loop feedback can suppress low-evidence extractions before they reach the enrichment stage, trading a small amount of recall for a substantial gain in precision.

Obtaining real-world trajectory: As discussed in Section IV, the enriched trajectories are synthesized from origin–destination pairs anchored on real incidents, since individual, real-time vehicular traces are neither publicly available nor free of privacy constraints. While this design is sufficient to validate the enrichment mechanism, the trajectories reflect plausible rather than observed travel demand. Future work will integrate real traces so that the enrichment can be assessed against genuine mobility patterns and at a scale beyond the present proof of concept.

Operational deployment and latency: The reported median latency of $25.5\,\mathrm{s}$ was measured under a provider's free-tier quota and therefore reflects rate-limit back-off rather than raw inference time. The architecture already supports both cloud-hosted and locally quantized deployment, so migrating to an unconstrained or more capable model is expected to reduce this figure substantially and to enable the accurate, continuous, and near-real-time regime for which the pipeline was proposed. Finally, broadening the set of citizen-generated sources beyond a single social-media API and adding more radio and news feed sources are concrete steps toward a more robust and representative event stream.

VII. Conclusion

This work investigated whether coupling foundation-model reasoning with a structured multi-agent pipeline can deliver continuous, explainable semantic enrichment of urban trajectories from real-time, citizen-generated data, an approach that existing enrichment methods, focused on historical data and static geographic primitives, do not address. We proposed and instantiated a model-agnostic, mixture-of-experts agentic system in which an orchestrator coordinates specialized collectors for traffic incidents, social media, news feeds and live radio, all communicating through a single versioned Pydantic contract and delegating natural-language understanding to a dedicated reasoning agent. The reasoner aligns the extracted events to vehicular trajectories through a spatio-temporal join and, operating under the ReAct paradigm with a human-in-the-loop protocol, produces enriched trajectories accompanied by auditable, natural-language justifications.

The system was instantiated for Belo Horizonte, Brazil, using open and citizen-generated data, and evaluated per agent and end to end. Our results show a clear contrast between the deterministic traffic backbone, which contributes a dense, fully schema-valid and georeferenced event stream, and the LLM-based agents, which are highly reliable in structural validity and groundedness, with at most 5.1% hallucination, while the radio agent remains limited by ASR-driven over-extraction. Crucially, the reasoning layer sustains this behaviour at low cost and under tight resource constraints, demonstrating that the architecture is economically viable even on a lightweight model.

Taken together, these findings support the central claim of the paper that a mixture-of-experts multi-agent design can transform heterogeneous, unstructured urban signals into structurally valid, well-grounded and auditable events at low cost. As a proof of concept, the system establishes a practical and extensible foundation for context-aware, explainable mobility services, and the directions outlined in Section VI chart a concrete path toward unsupervised, real-time operation.

References

[1] S. Chen, Y. Liu, W. Han, W. Zhang, and T. Liu, "A survey on llm-based multi-agent system: Recent advances and new frontiers in application," 2025. [Online]. Available: https://arxiv.org/abs/2412.17481

[2] W. Cao, Z. Wu, D. Wang, J. Li, and H. Wu, "Automatic user identification method across heterogeneous mobility data sources," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 2016, pp. 978–989.

[3] Y. Chen, Y. Tao, Y. Jiang, S. Liu, H. Yu, and G. Cong, "Enhancing large language models for mobility analytics with semantic location tokenization," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 262–273. [Online]. Available: https://doi.org/10.1145/3711896.3736937

[4] F. Meng, Y. Yuan, J. Ding, J. Feng, C. Han, and Y. Li, "Movefm-r: Advancing mobility foundation models via language-driven semantic reasoning," 2026. [Online]. Available: https://arxiv.org/abs/2509.22403

[5] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[6] R.-P. Schäfer, K.-U. Thiessenhusen, E. Brockfeld, and P. Wagner, "A traffic information system by means of real-time floating-car data," vol. 11, 01 2002.

[7] G. Chen, J. Li, J. Lu, and J. Zhou, "Human trajectory prediction via counterfactual analysis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9804–9813.

[8] N. Louzi, A. M. Arabiat, and M. AlJamal, "A counterfactual ai-based system for spatio-temporal traffic risk prediction and intelligent safety intervention in smart transportation systems," *Infrastructures*, vol. 11, no. 5, 2026. [Online]. Available: https://www.mdpi.com/2412-3811/11/5/152

[9] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, "Semantic trajectories modeling and analysis," *ACM Comput. Surv.*, vol. 45, no. 4, Aug. 2013. [Online]. Available: https://doi.org/10.1145/2501654.2501656

[10] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 336–343. [Online]. Available: https://doi.org/10.1145/1653771.1653818

[11] J. N. Vidal-Filho, V. C. Times, J. Lisboa-Filho, and C. Renso, "Towards the semantic enrichment of trajectories using spatial data infrastructures," *ISPRS International Journal of Geo-Information*, vol. 10, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2220-9964/10/12/825

[12] C. Gao, Z. Zhang, C. Huang, H. Yin, Q. Yang, and J. Shao, "Semantic trajectory representation and retrieval via hierarchical embedding," *Information Sciences*, vol. 538, pp. 176–192, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520305260

[13] J. N. Vidal-Filho, V. C. Times, J. Lisboa-Filho, and C. Renso, "Towards the semantic enrichment of trajectories using spatial data infrastructures," *ISPRS International Journal of Geo-Information*, vol. 10, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2220-9964/10/12/825

[14] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 02 2016.

[15] S. Hasan, C. Schneider, S. Ukkusuri, and M. C. Gonzalez, "Spatiotemporal patterns of urban human mobility," *Journal of Statistical Physics*, vol. 151, pp. 1–15, 04 2012.

[16] A. Prelipcean, G. Gidófalvi, and Y. Susilo, "Transportation mode detection – an in-depth review of applicability and reliability," *Transport Reviews*, pp. 1–23, 10 2016.

[17] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate gps trajectories," 11 2009, pp. 352–361.

[18] D. N. Campo, J. Conde, Álvaro Alonso, G. Huecas, J. Salvachúa, and P. Reviriego, "Real-time spatial retrieval augmented generation for urban environments," 2025. [Online]. Available: https://arxiv.org/abs/2505.02271

[19] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, May 2015. [Online]. Available: https://doi.org/10.1145/2743025

[20] Wikipédia, "Rss — wikipédia, a enciclopédia livre," 2025, [Online; accessed 19-novembro-2025]. [Online]. Available: \url{https://pt.wikipedia.org/w/index.php?title=RSS&oldid=71223166}