WISE-MAS: A Multi-Agent LLM Framework for Cultural Mobility Routing in Urban Environments

Carla B. Ferreira*
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
carlaferreira@dcc.ufmg.br

Selene M. Andrade*
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
selenemelo@dcc.ufmg.br

Antonio A. F. Loureiro
Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
loureiro@dcc.ufmg.br

***Abstract*—We present WISE-MAS, a generalist Multi-Agent System (MAS) framework for intelligent, interactive cultural mobility routing in urban environments. WISE-MAS combines Large Language Models (LLMs) with open geospatial data to guide users through a city's cultural Points of Interest (POIs) via natural-language conversation, routing them on foot or by public transit. The framework adopts a Mixture-of-Experts (MoE) architecture in which a conversational Orchestrator Agent dynamically coordinates two specialist agents: a Cultural Agent that retrieves and describes POIs through semantic search, and a Routing Agent that geocodes user locations and computes optimal pedestrian paths using Dijkstra's algorithm over the OpenStreetMap walking graph, and suggests direct public-transit connections from a GTFS stop-to-line feed. Human-in-the-Loop (HITL) interaction is embedded at each decision point to mitigate hallucinations and promote user agency. WISE-MAS runs entirely on open data and is designed to generalize to other cities with open geospatial and transit data; here, it is instantiated as a case study for Belo Horizonte, Brazil, using open municipal data. Evaluation combines human annotation, an LLM-as-a-Judge panel of three models (Mixtral-8x7B, Llama3.1-8B, DeepSeek-R1), and deterministic route validation. Results demonstrate high response coherence and clarity, and robust graceful degradation when specialist agents become unavailable. WISE-MAS is a replicable blueprint for culturally aware urban mobility assistants, with strong potential for deployment in other cities that rely on public transportation and cultural POI open data rather than per-city training.**

***Index Terms*—Multi-Agent System, Large Language Models, Agentic AI, Cultural Tourism, Urban Mobility, Mixture of Experts, Human-in-the-Loop, LLM-as-a-Judge, OpenStreetMap, GTFS.**

I. Introduction

Intelligent urban mobility systems are increasingly required to support not only commuting, but also cultural exploration by guiding tourists and residents through a city's Points of Interest (POIs) via natural-language interaction while integrating real geospatial data and public transit information. This problem lies at the intersection of agentic AI, urban computing, and cultural access: it requires semantic understanding of user intent, geospatial grounding, and step-by-step navigation, tasks that challenge both traditional recommendation systems and single-model LLM approaches [1], [2].

*Authors equally contributed to this work

Cultural tourism is a significant activity worldwide. Studies show that many people visit museums primarily while traveling [3] and that cultural tourism is a substantial and growing segment of international travel [4]. Despite this demand, no generic open-source system currently exists to guide users interactively through a city's cultural landscape, combining natural-language understanding, open geospatial data, and public-transit routing in a single agentic pipeline. Existing solutions are either geographically fixed proprietary products, domain-specific recommendation engines without navigation capability, or multi-day itinerary planners that do not address real-time conversational mobility [1], [2].

This paper presents **WISE-MAS** (**W**alkable **I**nteractive **S**ystem for cultural **E**xploration via **M**ulti-**A**gent **S**ystems), a generalist MAS framework for cultural mobility routing built entirely on open data and open-source components. WISE-MAS decomposes the problem into two specialist subproblems (cultural POI retrieval and geospatial routing) coordinated by a conversational Orchestrator Agent under a Mixture-of-Experts (MoE) design. The framework is intentionally designed to generalize across cities: it depends only on internationally standardized data formats (OpenStreetMap for road networks, GTFS for public transit, open municipal POI datasets) or readily available data (geocoding via Nominatim). We instantiate and evaluate WISE-MAS as a case study for Belo Horizonte, Brazil, a 2.4-million Latin American city with rich cultural infrastructure and publicly available open data, using its municipal open-data portal (BHTrans) and the GTFS transit feed.

The key contributions of this work are:

- A **generalist hierarchical MoE MAS architecture** that couples natural-language understanding with deterministic shortest-path computation, designed to be deployable in any city providing open geospatial and transit data;
- An **embedding-based semantic POI retrieval pipeline** over open municipal cultural datasets, with orthographic fuzzy matching as a fallback for robustness to user input errors;
- A **three-tier evaluation methodology** combining human annotation, a multi-model LLM-as-a-Judge panel, and deterministic route validation, applied to the Belo Horizonte case study;

979-8-3195-0167-7/26/$31.00 ©2026 IEEE

- An **analysis of fault-tolerant behavior** in conversational MAS, specifically the graceful degradation mechanism that continues user interaction when individual specialist agents fail.

The remainder of this paper is organized as follows: Section II presents background concepts; Section III discusses the related work; Section IV describes the WISE-MAS architecture; Section V describes the Belo Horizonte case study and open-data configuration; SectionVI reports the evaluation; and Section VII presents the concluding remarks and future work.

II. Background

A. Agents and Multi-Agent Systems

A computational agent is an autonomous entity that perceives its environment and acts to achieve defined objectives [5]. Multi-Agent Systems (MAS) combine multiple agents, each with a specific responsibility, whose coordinated behavior is more robust than any single agent acting in isolation [6]. In collaborative, hierarchical MAS, a coordinator agent integrates specialist outputs and manages the overall workflow. This architecture has been applied successfully to complex urban tasks [7].

B. Mixture of Experts (MoE) for Agentic Systems

Mixture of Experts [7] is an organizational strategy for MAS in which distinct, specialized agents handle different subsets of tasks. A routing mechanism (the Orchestrator Agent in this case) maps each incoming request to the most appropriate specialist via a function $r : \tau \mapsto a$, where τ is a task and a is the selected agent. This promotes modularity, reduces per-agent complexity, and is well-aligned with the ReAct (Reason-and-Act) paradigm of iterative tool use [8].

C. Human-in-the-Loop (HITL)

The HITL paradigm embeds human participation at key decision points in an agentic workflow to ensure accuracy, safety, and ethical accountability [9]. Rather than operating fully autonomously, the system returns control to the user at each decision step, soliciting location, confirming destination choices, and presenting route options for approval. This design is particularly important in mobility contexts, where misinformation about routes or destinations can have real-world safety consequences.

D. LLM-as-a-Judge Evaluation

LLM-as-a-Judge (LLMaaJ) [10] employs a language model as an automated evaluator of system outputs, assigning scores along defined quality dimensions (relevance, coherence, completeness, clarity, faithfulness). LLMaaJ approximates human judgment at scale for criteria that are difficult to formalize deterministically. Gu et al. [10] provide a comprehensive taxonomy of LLMaaJ methodologies, while Zheng et al. [11] demonstrate that inter-judge variability motivates using multi-model panels rather than relying on a single evaluator.

E. Shortest Path and Urban Routing

Route computation in urban networks is modeled as a shortest-path problem on a weighted directed graph, where intersections are vertices and road segments are edges weighted by length in meters. We use Dijkstra's algorithm [12], which is optimal for non-negative edge weights. Applied to the road graph of any city obtained from OSM via OSMnx [13], Dijkstra runs in milliseconds for walking routes bounded by typical time thresholds. Notice that other criteria, such as personal interests and city-related metrics (e.g., safe neighborhoods), could be used rather than a shortest-path solution.

III. Related Work

A. Urban Mobility Planning with LLM Agents

The most closely related line of work applies LLM-based agents to travel and mobility planning in urban environments. TravelPlanner [1] is the primary benchmark in this space, providing a sandbox with nearly 4 million data records accessible via 6 tools for evaluating multi-day, multi-constraint itinerary construction using language agents. It reveals that even GPT-4 achieves a final plan feasibility rate below 5%, motivating architectures that combine LLM reasoning with deterministic solvers.

ItiNera [2] is the most structurally similar system to WISE-MAS. It introduces Open-domain Urban Itinerary Planning (OUIP), which decomposes user requests, selects candidate POIs, and orders them via cluster-aware spatial optimization. WISE-MAS shares the POI-selection-then-routing decomposition, but differs in three key respects: (i) it integrates real-time conversational HITL at each planning step; (ii) it grounds srouting in a deterministic graph shortest-path solver over OSM rather than a spatial optimizer working on abstract coordinates; and (iii) it explicitly incorporates public-transit routing via GTFS alongside pedestrian routing. TravelAgent [14] similarly provides a personalized multi-agent travel assistant but relies on commercial APIs, whereas WISE-MAS is built entirely on open data sources.

Bhandari et al. [15] evaluate LLMs directly for urban mobility assessment tasks, finding that LLMs can reason about spatial patterns but require structured tool access for reliable numerical computation, a finding that directly motivates WISE-MAS's design choice of delegating all distance and route calculations to Dijkstra rather than to the LLM. The Constraint-Aware Route Recommendation framework [16] demonstrates that hierarchical LLM agents can extract and satisfy implicit routing constraints from natural language, improving route fidelity over single-model baselines; WISE-MAS adopts a similar hierarchical structure but grounds constraint resolution in its deterministic Routing Agent.

B. LLMs and Public Transit

TransitGPT [17] demonstrates that LLMs can significantly enhance the accessibility and usability of GTFS data for transit queries by using code generation to answer natural-language questions about transit schedules and routes, without

requiring users to understand the GTFS specification. WISE-MAS extends this vision by integrating GTFS-based transit routing as one layer within a broader multi-agent cultural mobility system, where transit suggestions emerge from a pre-processed stop-to-lines index rather than online GTFS parsing, a design that eliminates runtime API dependencies while maintaining routing accuracy.

C. LLMs for Urban Computing and Smart Cities

Ma et al. [4] present a mobility AI agent network work for intelligent transportation that connects people, vehicles, and road networks, using agents as behavioral simulation entities rather than interactive assistants. Bosse and Engel [18] combine virtual simulation with real-time data collected by chatbot agents for smart city applications, such as urban lighting and traffic management. Both differ from WISE-MAS in that they focus on infrastructure management rather than on end-user cultural navigation. Recent surveys on LLM-powered urban agents [19] provide a broad taxonomy of how LLMs interact with geospatial, temporal, trajectory, and textual inputs in urban settings. WISE-MAS fits within the geospatial-input and text-input categories of this taxonomy,combining OSM road network data, GTFS transit data, and POI coordinates with natural-language user interaction.

D. Agent-Based Cultural Heritage Systems

Amato et al. [20] propose an agent-based system for routing visitors to cultural sites within museums, treating visit scheduling as an accessibility problem. Vosinakis et al. [21] present multi-agent platforms for the dissemination of intangible cultural heritage in virtual environments. Rogers et al. [22] study driver route preference prediction, revealing that user preferences involve trade-offs (e.g., preferring longer routes if they reduce the number of turns) that are relevant to personalized navigation. WISE-MAS addresses urban-scale cultural access rather than intra-site routing, and incorporates user preference through conversational HITL rather than learned preference profiles.

E. LLM-as-a-Judge Evaluation

Gu et al. [10] provide a formal taxonomy of LLMaaJ systems and advocate for multi-judge panels to reduce inter-model variability. Zheng et al. [11] empirically establish that judge models differ systematically in leniency and scoring patterns, justifying our use of three architecturally distinct judges. Wang et al. [23] propose the Self-Instruct framework for generating evaluation rubrics, which informs our human-annotation criteria. Masterman et al. [24] survey emerging agentic AI architectures and evaluation practices, situating our evaluation design within the broader landscape of agenticsystem assessment.

F. Positioning of WISE-MAS

Table I summarizes how WISE-MAS differs from the most closely related systems. The key differentiator is the combination of: real-time conversational HITL, deterministic graph routing over open OSM data, standardized GTFS transit integration, semantic embedding-based POI retrieval, and a design that generalizes across any city, providing open geospatial infrastructure.

TABLE I
COMPARISON OF RELATED SYSTEMS

System	Conv. MAS	HITL	OSM routing	GTFS transit	Open data
TravelPlanner [1]	✓	–	–	–	–
ItiNera [2]	–	–	–	–	partial
TravelAgent [14]	✓	–	–	–	–
TransitGPT [17]	✓	–	–	✓	partial
Amato et al. [20]	✓	–	–	–	–
WISE-MAS (ours)	✓	✓	✓	✓	✓

IV. WISE-MAS FRAMEWORK

WISE-MAS implements a hierarchical MoE MAS with three agents: an Orchestrator Agent, a Cultural Agent, and a Routing Agent. Figure 1 illustrates the proposed architecture. All agents use the Smolagents framework [25] from Hugging Face and follow the ReAct paradigm [8]: at each step, the agent produces a reasoning trace, generates a code block or tool call, observes the result, and iterates. The LLM backbone is Qwen3 [26], served locally via Ollama through the LiteLLM proxy, enabling fully offline operation without dependence on external proprietary APIs.

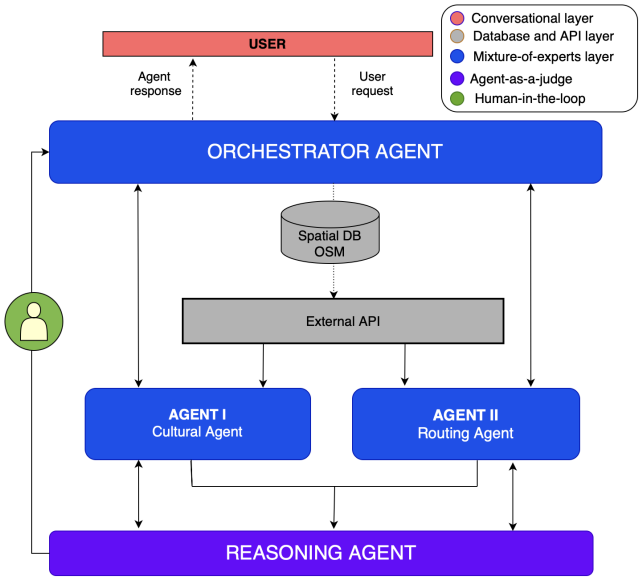


Fig. 1. WISE-MAS architecture. The Orchestrator Agent coordinates the Cultural Agent (Agent I) and Routing Agent (Agent II) over a conversational layer. Colors indicate system layers: conversational (red), database/API (gray), MoE (blue), judge (purple), HITL (green).

Table II summarizes the system specification according to the agent taxonomy of [9].

A. Orchestrator Agent

The Orchestrator Agent is the system entry point. It has no domain tools of its own; instead, it coordinates and invokes the

TABLE II
WISE-MAS SYSTEM SPECIFICATION

Dimension	Characteristic
Control architecture	LLM as deliberative component (orchestrator); hybrid (specialists)
Learning	Fixed weights; no fine-tuning or RL at runtime
World model	Model-based: OSM road graph, GTFS transit data, geocoded POI coordinates
Observability	Partially observable: text descriptions, static datasets, no real-time sensors
Social context	Collaborative multi-agent
Abstraction level	Hierarchical: coordinator commands specialist agents
Cognitive architecture	Cognitive: multi-turn NL understanding, dynamic agent routing, code-as-plan generation
Specialization	Domain-specific: cultural mobility routing

two specialists as high-level callable functions and decides at each conversational step which agent to activate. This decision implements the MoE routing function:

$$r(\tau) = \begin{cases} \text{Routing Agent} & \text{if } \tau \in \{\text{proximity, route, navigation}\} \\ \text{Cultural Agent} & \text{if } \tau \in \{\text{description, search, category}\} \end{cases} \quad (1)$$

The orchestrator maintains multi-turn conversational memory so that the user's location and destination choices persist across messages. Since each specialist is executed as a stateless sub-process with no memory of prior interactions, the orchestrator re-injects origin and destination into every specialist call.

A set of explicit constraints encoded in the orchestrator's system prompt (Table III) ensures factual grounding and prevents hallucinated POIs, distances, or transit lines. In particular, the "factual anchoring" constraint prohibits the orchestrator from inventing any location, address, distance, or bus line: it may only relay facts returned by specialist tool calls. This design directly addresses the hallucination problem identified in [15] for LLMs performing urban mobility tasks.

TABLE III
ORCHESTRATOR CONSTRAINTS AND MOTIVATIONS

Constraint	Motivation
Intent-based routing	Route/proximity queries to Routing Agent; description/search queries to Cultural Agent, preventing geographic errors.
One specialist per step	Never invoke both agents in the same code block; prevents blind action and timeout violations.
Context relay	Provide origin and destination in every specialist call, compensating for stateless sub-execution.
Factual anchoring	Never fabricate POIs, addresses, distances, or transit lines; use only tool-returned data. Mitigates hallucination [15].
`final_answer` use	Always terminate via `final_answer()`; never shadow it with a variable.

The orchestrator also enforces graceful degradation: if a specialist becomes unresponsive or times out, the orchestrator detects the failure, issues a user-visible warning, and continues the interaction using only the remaining available agent, a critical property for real-world deployment in mobile environments with unreliable network connectivity.

B. Routing Agent

The Routing Agent localizes the user, identifies nearby cultural POIs, and computes walking or transit routes to a chosen destination. The workflow illustrated in Figure 2 proceeds through four sub-steps.

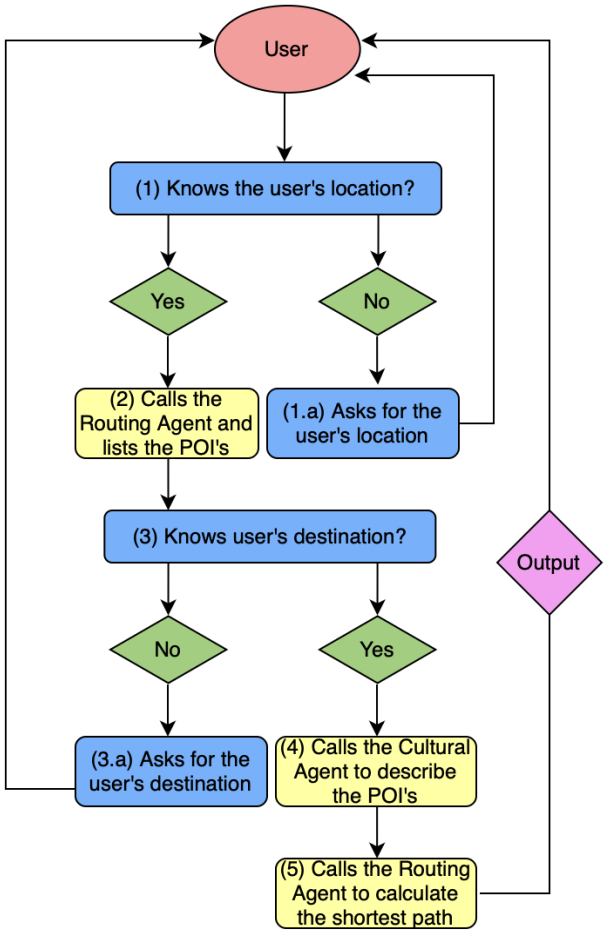


Fig. 2. Orchestrator decision flow. Yellow boxes represent specialist agent calls; blue boxes are orchestrator decision or query steps.

a) Geocoding: Free-text location descriptions are normalized and geocoded via Nominatim/OSM [27], constrained to the city's bounding box. Orthographic fuzzy matching resolves common spelling and accent errors. Results are cached for the duration of the conversation. If Nominatim fails, a fallback computes the centroid of the specified neighborhood to maximize the probability of locating nearby POIs.

b) Nearest-POI Search: POI coordinates (sourced from the city's open data portal) are indexed for spatial search. The k-Nearest Neighbor algorithm (k=5, Haversine distance) identifies the closest POIs to the geocoded origin. For queries

that name a specific destination, substring matching and textual similarity determine the target POI.

c) Pedestrian Routing: The city's road network is obtained from OpenStreetMap via OSMnx [13] and loaded into a weighted directed graph. Walking routes are computed using Dijkstra's algorithm [12] with edge weights equal to segment length in meters, a well-established approach for city-scale pedestrian navigation [13]. Consecutive road segments are aggregated into turn-by-turn instructions. A configurable walkability threshold (default: 30 minutes at 5 km/h) determines whether pedestrian routing is feasible; if not, use transit routing as the fallback.

d) Public Transit Routing: Bus and metro data are sourced from the GTFS standard [17], used internationally to represent public transit schedules and routes. Because the raw GTFS file can be too large for real-time parsing, WISE-MAS preprocesses it offline into a stop-to-lines index that maps each bus stop to the set of lines that serve it. At runtime, the agent queries nearby stops and returns direct transit options with suggested boarding and alighting stops. Metro or rail stations with stable schedules are encoded as static knowledge with embedded coordinates. This design eliminates runtime dependency on external transit APIs, making responses fast and resilient to network failures, an approach that builds on the code-as-agent paradigm demonstrated by TransitGPT [17] while optimizing specifically for low-latency conversational use.

C. Cultural Agent

The Cultural Agent retrieves and describes cultural POIs from the city's open data portal. It follows a three-tier retrieval hierarchy:

1) Exact matching: name, category, or neighborhood keyword search;
2) Category filtering: fixed category set (museum, theater, library, park, cultural center, etc.);
3) Semantic embedding search: cosine similarity over precomputed POI vectors as a fallback for contextual queries (e.g., "an open place with trees").

For semantic search, each POI in the dataset is pre-encoded offline once via `nomic-embed-text` and cached on disk. At query time, the user's request is encoded into the same embedding space and ranked by cosine similarity against all stored POI vectors. This enables the agent to match semantically related requests to categorically appropriate POIs without requiring exact keyword matches. For example, mapping "I would like to see some paintings" to art museums even when the user does not use the word "museum." This retrieval design is conceptually related to Retrieval-Augmented Generation (RAG) approaches applied to tourism recommendation [28], but grounds responses exclusively in tool-returned data rather than generating new content about POIs.

D. Human-in-the-Loop Interaction

WISE-MAS implements HITL at four decision points per conversational session: (1) soliciting the user's location; (2) presenting nearby POIs and asking which to learn more about; (3) displaying POI descriptions and asking which to route to; and (4) presenting the computed route for user confirmation. This design ensures that the final itinerary emerges from a sequence of human-confirmed choices, providing transparency and enabling users to correct agent misunderstandings before they propagate into routing errors.

E. City-Generalization Design

WISE-MAS is designed to generalize across cities. The framework confines its city-specific layer to four data inputs: (i) OSM road network data (globally available via OSMnx); (ii) Nominatim geocoding (globally available); (iii) GTFS transit feed (available for hundreds of cities internationally); (iv) an open-data POI dataset in any tabular format (available from most municipal open-data portals); and (v) a configured bounding box for Nominatim geocoding. No city-specific machine learning training is required. Adapting WISE-MAS to a new city requires only (a) downloading its OSM graph and GTFS feed, (b) importing its POI dataset, and (c) generating POI embeddings once offline. This architecture directly addresses the scalability limitation identified in the related work that requires per-city retraining [2], [15]. We stress that this is an argument from construction, since we instantiated and evaluated WISE-MAS in a single city. We therefore claim portability by design rather than demonstrated generalization, and identify validation across multiple cities as the primary open issue for future work. Two properties, however, define the scope within which the architecture is expected to transfer with minimal discrepancies between cities. First, the data interfaces are model-agnostic. This means GTFS can encode bus, metro, and rail under a single schema, so ingestion does not change as a city's model composition changes. Second, medium-to-large cities usually publish both GTFS feeds and open cultural datasets through municipal portals, expanding the number of cities the framework is expected to transfer to and that our future multi-city validation should sample. Extending beyond our use case requires verifying data availability and catalog structures under different city regimes. Table IV surveys GTFS availability per transport mode in five reference cities.

TABLE IV
AVAILABILITY OF OPEN GTFS FEEDS BY TRANSPORT MODE

City	**Transport Mode**			
	Bus	***Metro***	***Tram***	***Ferry***
Belo Horizonte, BR [29]	Yes[a]	No[b]	–[c]	–
São Paulo, BR [30]	Yes	Yes	–	–
Rio de Janeiro, BR [31]	Yes	No	–	–
Paris, FR [32]	Yes	Yes	Yes	–
Amsterdam, NL [33]	Yes	Yes	Yes	Yes

[a]Mode is covered by a publicly downloadable GTFS feed.
[b]Mode operates in the city but is not covered by any open feed.
[c]Mode does not exist in the city.

TABLE V
SPECIALIST AGENT TOOLS

Agent	Tool	Description
Routing	`localize_user`	Geocodes a free-text location to (lat, lon).
	`nearby_pois`	Lists the k nearest cultural POIs.
	`find_poi`	Finds a specific POI by name.
	`nearest_poi`	Returns the single nearest POI.
	`walk_route`	Computes the shortest walking route via Dijkstra.
	`transit`	Suggests transit options via GTFS index.
Cultural	`search_pois`	Free-text search over the open POI dataset.
	`filter_by_type`	Filters POIs by category.
	`semantic_search`	Cosine-similarity search over embeddings.

V. CASE STUDY: BELO HORIZONTE, BRAZIL

Belo Horizonte (BH), the capital of Minas Gerais state, Brazil, is a mid-sized Latin American city (approximately 2.4 million inhabitants) with a rich cultural infrastructure: it hosts over 40 museums, 30 theaters, numerous cultural centers, and the CCBB (Bank of Brazil Cultural Center), ranked among the 50 most visited cultural institutions in the world [34]. The city makes its cultural data publicly available through the PBH Open Data Portal, providing regularly updated files of cultural venues and tourist attractions with category labels, addresses, and geolocation [35]. The municipal transit authority BHTrans publishes GTFS feeds covering the city's entire bus network, and OSM provides a high-quality road graph. This combination of cultural richness and open data availability makes BH an ideal case study for WISE-MAS.

Table V lists all tools exposed by each agent.

A. Data Configuration for Belo Horizonte

a) POI Dataset: The PBH portal provides two CSV files: `pontos_cultura.csv` (cultural venues including museums, theaters, libraries, cultural centers) and `atrativo_turistico.csv` (tourist attractions), with five category labels: `NATURAL`, `GASTRONOMY AND SHOPPING`, `TECHNICAL AND SCIENTIFIC ACHIEVEMENTS`, `CULTURAL EVENTS`, and `ENTERTAINMENT AND LEISURE`.

b) Road Network: The OSM road graph of Belo Horizonte is downloaded via OSMnx, yielding a graph of moderate size (order of 10^5 nodes) for which Dijkstra routing runs in milliseconds for walking-range queries.

c) Transit Data: The BHTrans GTFS feed covers the full municipal bus network. The offline preprocessor builds a stop-to-lines index from the GTFS `stop_times` and `trips` files. Metro stations (a smaller, stable network of 19 stations) are encoded as static knowledge with hardcoded coordinates and line labels to avoid dependence on runtime GTFS parsing.

d) Geocoding Bounding Box: Nominatim queries are constrained to the geographic bounding box of Belo Horizonte to prevent geocoding results from other cities with similar street names.

B. Implementation Notes

All agents are implemented in Python using Smolagents 1.x. The embedding model (`nomic-embed-text`) runs locally via Ollama. Geocoding cache uses an in-memory dictionary keyed by a normalized location string. The GTFS preprocessor runs once at system startup and caches results to disk; subsequent runs load the index directly. The full system, including models, runs on a commodity laptop without GPU acceleration, demonstrating that WISE-MAS is deployable without specialized hardware.

VI. EVALUATION

A. Evaluation Design

We evaluate the Belo Horizonte instantiation of WISE-MAS through a three-tier methodology: (1) **Human annotation** by two independent raters; (2) **LLM-as-a-Judge** scoring by three distinct judge models; and (3) **Deterministic route validation** against Google Maps reference data. The three tiers are complementary: human annotation captures nuanced judgment; LLMaaJ provides scalable multi-dimensional scoring; and deterministic validation anchors the evaluation in factually verifiable ground truth. This multi-tier design follows the hybrid evaluation pattern advocated by Masterman et al. [24] for agentic AI systems.

B. Human Annotation

Using `qwen3:latest` for response generation, we constructed 17 test cases: 12 POI retrieval queries using semantically related terms not present verbatim in the PBH dataset (all departing from Station Square, a central landmark), and 5 routing pairs with known ground-truth routes verified against Google Maps. Two independent raters (the authors, who are familiar with the city and its locations, being capable of a high level of critical analysis regarding relevance) scored each response on the 0–3 scale of Wang et al. [23]: **3** = valid and satisfactory; **2** = acceptable with minor errors; **1** = relevant but with significant content errors; **0** = irrelevant or invalid.

TABLE VI
HUMAN EVALUATION: POI SEMANTIC RETRIEVAL (12 QUERIES, DEPARTING FROM PRAÇA DA ESTAÇÃO)

Query	R1	R2	Total
"open place with trees"	3	3	6
"I want to watch a play"	2	1	3
"painting or sculpture exhibition"	1	1	2
"typical local cuisine"	3	3	6
"nice view of the city"	0	0	0
"I want to watch football"	0	0	0
"street food"	0	0	0
"planetarium or science museum"	3	2	5
"walk surrounded by greenery"	1	2	3
"I'd like to see some paintings"	3	3	6
"bowling"	0	0	0
"skydiving"	0	0	0
Mean (out of 6)			**2.58**

Routing responses (mean 4.60/6) significantly outperform POI retrieval responses (mean 2.58/6), confirming the advantage of deterministic graph computation for precision-critical

TABLE VII
HUMAN EVALUATION: ROUTE COMPUTATION (5 PAIRS)

Origin	Destination	R1	R2	Total
Praça da Estação	Mercado Novo	3	3	6
Praça Rui Barbosa	Museu de Artes e Ofícios	2	2	4
Praça Floriano Peixoto	Cine Santa Tereza	1	1	2
Praça Raul Soares	Museu da Imagem e do Som	3	3	6
Praça Sete de Setembro	Biblioteca Pública Infantil	2	3	5
Mean (out of 6)				**4.60**

tasks, which are consistent with the finding of Handari et al. [15] that LLMs alone are unreliable for tasks requiring numerical precision.

C. LLM-as-a-Judge Evaluation

We assembled a panel of three judge models with distinct architectures, following the multi-judge protocol recommended by Gu et al. [10] to reduce individual judge bias: **Mixtral-8x7B** [36] (sparse MoE), **Llama3.1-8B** [37] (dense transformer), and **DeepSeek-R1** (chain-of-thought reasoning specialist). Each judge evaluated 10 end-to-end conversational scenarios on five criteria (1–5 scale): Relevance, Coherence, Completeness, Clarity, and Faithfulness.

TABLE VIII
LLM-AS-A-JUDGE PANEL: MEAN SCORES (1–5) OVER 10 SCENARIOS

Criterion	Mixtral-8x7B	Llama3.1-8B	DeepSeek-R1
Relevance	3.90	4.60	3.60
Coherence	4.40	3.50	4.10
Completeness	3.80	2.70	2.60
Clarity	4.80	4.10	3.70
Faithfulness	3.90	3.00	3.60

All three judges rate the system above 3.5 on Relevance and Clarity, confirming that responses are appropriate and understandable. Completeness and Faithfulness receive lower, more variable scores, indicating that some responses omit expected steps (e.g., not presenting the full route when the user has not yet confirmed the destination) or contain minor factual divergences. Inter-judge agreement is 72%, corroborating Zheng et al.'s [11] finding that judge models differ systematically in leniency: Mixtral is the most lenient, Llama3.1 and DeepSeek-R1 the most critical.

D. Results Discussion

Three principal findings emerge from the evaluation.

a) Routing is more reliable than semantic retrieval: The gap between routing scores (4.60/6) and POI retrieval scores (2.58/6) is substantial. Failure cases in POI retrieval cluster around two patterns: (i) queries that map to activities absent from the PBH dataset (e.g., "street food", "skydiving", "bowling"), where the system falls back to proximity-based suggestions without informing the user; and (ii) KNN results limited to k=5 with origin-centroid geocoding, which may miss more relevant POIs at greater distances when the user's location is imprecisely specified. This confirms that the Routing Agent's deterministic design is more robust than the Cultural Agent's LLM-driven retrieval for precision tasks.

b) Multi-judge evaluation is necessary: The 72% inter-judge agreement and the score variability across judges (e.g., Completeness ranges from 2.60 to 3.80) confirm that single-judge evaluation provides unreliable quality estimates [10]. The three-model panel identifies consistent strengths (Clarity, Relevance) and consistent weaknesses (Completeness, Faithfulness) while filtering out idiosyncratic per-model biases.

c) Graceful degradation is effective: In test scenarios where one specialist agent was disabled, the orchestrator successfully detected the failure, issued an informative warning to the user, and continued the conversation with the remaining available agent. This fault-tolerant behavior, implemented through the orchestrator's observation-retry mechanism, sustains user interaction without system crash, which is a critical property for mobile deployment in environments with variable network availability.

VII. CONCLUSION AND FUTURE WORK

This paper presented WISE-MAS, a generalist Multi-Agent LLM framework for cultural mobility routing in urban environments, instantiated as a case study for Belo Horizonte, Brazil. WISE-MAS integrates semantic POI retrieval over open municipal data, deterministic shortest-path pedestrian routing over OSM, GTFS-based public-transit suggestions, and conversational HITL within a hierarchical Mixture-of-Experts MAS orchestrated by a ReAct-style LLM agent. Evaluation with a three-model LLMaaJ panel and human annotation confirms high routing fidelity and response coherence, while identifying semantic retrieval completeness and response faithfulness as the primary targets for improvement. The framework is designed to generalize to any city that provides open geospatial and transit data, requiring no per-city machine-learning retraining.

Future directions include:

- Multi-city validation: instantiating WISE-MAS in additional cities with different urban morphologies (e.g., European grid cities, Asian megacities) to verify the generalization claim empirically.
- Improved semantic retrieval: adopting hybrid distance-and-similarity ranking, larger k, and richer POI datasets to address the coverage limitations identified in the BH evaluation.
- Real-time data integration: incorporating live bus position APIs (where available), real-time event schedules, and venue opening hours to increase temporal situational awareness.
- Personalization: adding a user-profile agent that persists preference history across sessions to enable proactive destination suggestions.
- Multimodal and accessibility extensions: voice-to-text input, map-based route visualization, and wheelchair-accessible routing variants.

- Scalable evaluation: constructing a larger city-walking benchmark inspired by TravelPlanner [1] for systematic comparison with related systems across multiple cities.
- Improvement to the ablation study: internal comparisons across different versions, including a single-agent approach and evaluation with new metrics.

ACKNOWLEDGMENT

This research was partially sponsored by CNPq grant 405940/2022-0, and FAPESP grant 2023/00673-7.

REFERENCES

[1] J. Xie, K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su, "TravelPlanner: A benchmark for real-world planning with language agents," in *Forty-First International Conference on Machine Learning (ICML)*, 2024.

[2] Y. Tang, Z. Wang, A. Qu, Y. Yan, Z. Wu, D. Zhuang, J. Kai, K. Hou, X. Guo, J. Zhao *et al.*, "ItiNera: Integrating spatial optimization with large language models for open-domain urban itinerary planning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 1413–1432.

[3] M. de Sólon Gosling *et al.*, "Vamos fazer algo diferente? um estudo exploratório sobre motivações de visitação a museus," p. 25, 2017.

[4] H. Ma, Y. Liu, Q. Jiang, B. Y. He, X. Liao, and J. Ma, "Mobility AI agents and networks," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5124–5129, 2024.

[5] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2020.

[6] N. R. Jennings, "On agent-based software engineering," *Artificial Intelligence*, vol. 117, no. 2, pp. 277–296, 2000.

[7] S. Chen, Y. Liu, W. Han, W. Zhang, and T. Liu, "A survey on LLM-based multi-agent systems: Recent advances and new frontiers in application," *arXiv preprint arXiv:2412.17481*, 2025.

[8] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[9] W. Reiznautt, E. Cerqueira, D. Cunha, L. Villas, A. Loureiro, D. Rosario, A. Souza, and N. Fonseca, *Agentes Inteligentes para Configuração de Redes de Computadores: Da Teoria à Prática com LLM, SLM, RAG e IA Agêntica*, May 2026, pp. 105–159.

[10] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on LLM-as-a-judge," *The Innovation*, vol. 7, no. 6, 2026.

[11] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] E. W. Dijkstra, "A note on two problems in connexion with graphs," pp. 269–271, 1959.

[13] G. Boeing, "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[14] A. Chen, X. Ge, Z. Fu, Y. Xiao, and J. Chen, "TravelAgent: An AI assistant for personalized travel planning," in *arXiv preprint arXiv:2409.08069*, 2024.

[15] P. Bhandari, A. Anastasopoulos, and D. Pfoser, "Urban mobility assessment using LLMs," in *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, 2024, pp. 67–79.

[16] J. Jain, V. Bagadia, S. Manchanda, and S. Ranu, "Constraint-aware route recommendation from natural language via hierarchical LLM agents," in *arXiv preprint arXiv:2510.06078*, 2024.

[17] S. Devunuri and L. J. Lehe, "TransitGPT: A generative AI-based framework for interacting with GTFS data using large language models," *Public Transport*, vol. 17, pp. 319–345, 2025.

[18] S. Bosse and U. Engel, "Real-time human-in-the-loop simulation with mobile agents, chat bots, and crowd sensing for smart cities," *Sensors*, vol. 19, no. 20, p. 4356, 2019.

[19] J. Yan *et al.*, "Large language model powered intelligent urban agents: Concepts, capabilities, and applications," 2025, arXiv:2507.00914.

[20] F. Amato, F. Moscato, V. Moscato, F. Pascale, and A. Picariello, "An agent-based approach for recommending cultural tours," *Pattern Recognition Letters*, vol. 131, pp. 341–347, 2020.

[21] S. Vosinakis, N. Avradinis, and P. Koutsabasis, "Dissemination of intangible cultural heritage using a multi-agent virtual world," in *Advances in Digital Cultural Heritage: International Workshop*. Springer, 2018, pp. 197–207.

[22] S. Rogers, C.-N. Fiechter, and P. Langley, "An adaptive interactive agent for route advice," in *Proceedings of the Third Annual Conference on Autonomous Agents*, 1999, pp. 198–205.

[23] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 13 484–13 508.

[24] T. Masterman, S. Besen, M. Sawtell, and A. Chao, "The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey," in *arXiv preprint arXiv:2404.11584*, 2024.

[25] Hugging Face, "Smolagents: A lightweight library for building LLM agents," 2024. [Online]. Available: https://github.com/huggingface/smolagents

[26] A. Yang, A. Li, B. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[27] OpenStreetMap Foundation, "Nominatim: Geocoding for OpenStreetMap data," 2024. [Online]. Available: https://nominatim.openstreetmap.org

[28] A. Banerjee, A. Satish, and W. Wörndl, "Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation," in *International Workshop on Recommender Systems for Sustainability and Social Good*. Springer, 2024, pp. 19–34.

[29] Prefeitura de Belo Horizonte, "Portal de dados abertos: GTFS do sistema de transporte coletivo," https://dados.pbh.gov.br/, 2026, accessed: Sep. 8, 2026.

[30] São Paulo Transporte S.A. (SPTrans), "Portal do desenvolvedor: GTFS e API olho vivo," https://www.sptrans.com.br/desenvolvedores/, 2026, accessed: Sep. 8, 2026.

[31] Prefeitura da Cidade do Rio de Janeiro, "DATA.RIO: GTFS do sistema de transporte por Ônibus," https://www.data.rio/, 2026, accessed: Sep. 8, 2026.

[32] Île-de-France Mobilités, "Open data platform (PRIM): GTFS feeds for the île-de-france region," https://prim.iledefrance-mobilites.fr/, 2026, accessed: Sep. 8, 2026.

[33] NDOV Loket, "Dutch national public transport data: GTFS feeds (OVapi)," https://ndovloket.nl/, 2026, accessed: Sep. 8, 2026.

[34] Diário do Comércio, "Museus de belo horizonte entre os mais visitados do mundo," 2022. [Online]. Available: https://diariodocomercio.com.br/variedades/museus-belo-horizonte-mais-visitados-mundo/

[35] Prefeitura de Belo Horizonte, "Portal de dados abertos da PBH," 2024. [Online]. Available: https://dados.pbh.gov.br/dataset/

[36] Mistral AI, "Mixtral of experts," 2023. [Online]. Available: https://mistral.ai/news/mixtral-of-experts/

[37] Meta AI, "Llama 3: Meta's next generation open source language model," 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/