

# CityFlow: Mixture-of-Experts Multi-Agent Collaboration for Personalized Urban Route Planning

Qilong Wang  
Beijing Normal University  
13623753581@163.com

June 2, 2026

## Abstract

Personalized urban route planning requires reconciling heterogeneous user preferences—culinary exploration, scenic sightseeing, leisure pacing, or speed-run touring—with spatial, temporal, and budgetary constraints. Existing approaches either rely on classical optimization solvers that cannot interpret natural language intent, or single-LLM pipelines that hallucinate POIs and produce spatially incoherent routes. We present **CityFlow**, a Mixture-of-Experts (MoE) multi-agent framework that decomposes route planning across 11 domain-specific experts orchestrated by a LangGraph state machine. The system features: (1) scene-aware expert dispatch that classifies user intent into five scenario types and activates only relevant experts; (2) parallel wave-based expert execution across two dependency-ordered waves; (3) a review-rework feedback loop evaluating proposals against six quality dimensions; and (4) algorithmic time smoothing that validates and corrects LLM-generated temporal allocations. Evaluated on 5 core scenarios covering Zhuhai, China (2000+ POIs), CityFlow achieves an average score of 7.6/10 with 5/5 pass rate, with individual scenario scores ranging from 7 to 8. We document a systematic optimization journey from a baseline of 65.2/100 to 81.2/100 across 100 test scenarios, identifying seven categories of failure modes and their resolutions. The system delivers first results within 5 seconds via Server-Sent Events streaming.

## 1 Introduction

Urban route planning is a constrained combinatorial optimization problem. A user seeking “a one-day seafood food tour in Zhuhai” expects a route that balances restaurant diversity, geographic proximity, time feasibility, and experiential richness—requirements that are difficult to formalize in a traditional constraint specification.

Classical approaches formulate this as a Time-Side Window Traveling Problem (TSPTW) [2], where a set of points of interest (POIs) must be visited within their operating hours while minimizing travel time. While exact solvers produce feasible routes, they require explicit constraint definitions and cannot interpret natural language preferences such as “romantic sunset dinner” or “energetic speed-run touring.”

Large Language Models (LLMs) excel at understanding user intent [3], but suffer from three critical limitations in spatial planning: (1) *hallucination*—recommending non-existent POIs [4]; (2) *spatial blindness*—inability to reason about geographic distances from textual descriptions alone; and (3) *optimism bias*—invariably attempting to satisfy impossible requests rather than declining gracefully.

We observe that human travel planning naturally decomposes into specialized concerns: a food enthusiast evaluates restaurants differently from a nature lover assessing hiking trails, and both differ from a logistics expert calculating inter-POI travel times. This suggests a **Mixture-of-Experts** (MoE) architecture [5] where domain-specific modules contribute proposals that are reviewed, refined, and synthesized into a coherent route.

## 1.1 Contributions

1. **MoE Multi-Agent Architecture:** We design an 11-expert system with scene-aware dispatch, where experts in scenic, culinary, cultural, natural, transportation, accommodation, and local-guide domains contribute proposals in parallel waves, orchestrated by a LangGraph state machine.
2. **Review-Rework Feedback Loop:** A quality assurance mechanism evaluates proposals against six dimensions (intent match, POI quality, geographic continuity, scene diversity, time feasibility, category match) and triggers targeted rework when scores fall below threshold.
3. **Prompt Engineering for Refusal:** We demonstrate that teaching LLMs to reject infeasible requests via explicit prompt instructions—rather than architectural components—is sufficient to handle impossible scenarios.
4. **Algorithmic Time Smoothing:** We introduce `_smooth_times()`, a post-processing layer that detects and corrects LLM-generated temporal anomalies (gaps, inversions, overflows).
5. **Detailed Optimization Case Study:** We document a systematic debugging journey from 65.2 to 81.2 points across 100 scenarios, identifying seven categories of failure modes and their resolutions.

## 2 Related Work

### 2.1 Classical Route Optimization

The Vehicle Routing Problem with Time Windows (VRPTW) and its single-vehicle variant TSPTW have been extensively studied [1]. Solomon’s benchmark [2] established the formulation, and modern solvers combine branch-and-bound with local search metaheuristics (2-opt, LKH). These methods produce provably feasible routes but require explicit constraint specification—they cannot process “I want a relaxing afternoon by the sea.”

### 2.2 LLM-Based Travel Planning

Recent work explores using LLMs for itinerary generation. TripPlanner [7] uses a single LLM to generate multi-day itineraries, but suffers from hallucinated POIs and poor spatial reasoning. TravelBot [8] incorporates retrieval-augmented generation (RAG) to ground recommendations in real POI databases. However, single-LLM approaches share a fundamental limitation: they tend to anchor on the most common interpretation of user input, producing homogeneous recommendations regardless of scenario type.

## 2.3 Multi-Agent Systems

The Mixture-of-Experts paradigm [5] has been adapted for LLM routing in models like Mixtral [6], where different experts handle different token distributions. We extend this concept to the *planning* domain, where experts represent domain knowledge (food, transportation, accommodation) rather than token distributions. General-purpose multi-agent frameworks like AutoGen [9] and CrewAI provide agent orchestration primitives, but lack domain-specific optimization for spatial planning tasks.

## 2.4 LLM Limitations in Planning

Prior work has documented LLM hallucination in knowledge-intensive tasks [4]. In the route planning context, we identify an additional failure mode: *optimism bias*—LLMs invariably attempt to satisfy user requests, even when the request is physically infeasible (e.g., visiting three islands in one day when ferry transit alone requires six hours). Our prompt-based refusal mechanism addresses this without dedicated architectural components.

# 3 System Architecture

## 3.1 Overview

CityFlow follows a five-phase pipeline architecture (Figure 1):

1. **Intent Parsing:** The *rule\_guard* module parses natural language input into a structured intent representation, loads candidate POIs from the database, and generates meta-rules for downstream constraints.
2. **Scene Classification:** The *expert\_router* classifies user intent into one of five scenario types using LLM-based classification with structured output, then computes expert activation weights.
3. **Parallel Expert Dispatch:** Domain experts execute in two waves—Wave 1 contains data-independent experts (POI, food, weather, destination) that run concurrently; Wave 2 contains dependency-aware experts (traffic, hotel, local guide) that require Wave 1 results.
4. **Review & Rework:** The *review* module evaluates proposals against six quality dimensions using both algorithmic checks and LLM judgment; if the score falls below threshold, targeted rework is triggered.
5. **Route Synthesis:** The *synthesizer* assembles the final route via TSPTW constraint satisfaction, followed by algorithmic time smoothing and narrative generation.

## 3.2 Scene-Aware Expert Dispatch

The *expert\_router* classifies user input into one of five scenario types using LLM-based classification with structured output. Table 1 shows the scenario types and their activated expert sets.

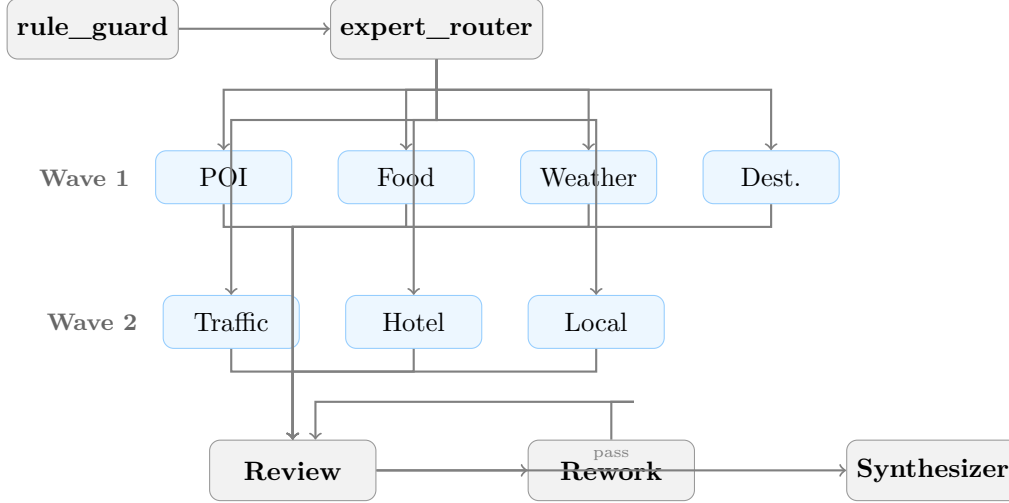


Figure 1: CityFlow MoE architecture. User input flows through rule\_guard and expert\_router, then dispatches to parallel expert waves. The review-rework loop iterates until quality threshold is met.

Table 1: Scenario types and their activated expert sets.

Scenario	Description	Primary Experts
Food Exploration	Culinary-focused	Food, Local, POI
Destination-Focused	Single major attraction	Destination, POI, Traffic
Speed-Run Touring	Dense checkpoint visiting	POI, Traffic, Hotel
Leisure	Slow-paced, few stops	POI, Weather, Local
Sightseeing	General tourism	POI, Food, Weather, Traffic

The router computes activation weights for each expert based on scenario relevance. For food exploration, the Food Expert receives weight 1.0 while the Hotel Expert receives 0.2; for destination-focused scenarios, the Destination Expert receives 1.0 with relaxed geographic constraints.

### 3.3 Expert Network

Table 2: Expert modules organized by execution wave.

Wave	Expert	Domain	Output
Wave 1	POI Expert	Scenic spots	Ranked candidate POIs by category
	Food Expert	Restaurants	Sub-categorized dining (5 sub-types)
	Weather Expert	Climate	Activity feasibility scores
	Destination Expert	Named locations	Core destination + radius POIs
Wave 2	Traffic Expert	Transportation	Inter-POI travel time estimates
	Hotel Expert	Accommodation	Rest stop suggestions
	Local Expert	Hidden gems	Non-mainstream high-rated places
Wave 3	Review	Quality control	6-dimension scoring + feedback
Wave 4	Synthesizer	Route assembly	TSPTW-optimized final route

Each expert operates independently with access to the shared POI database (2000+ entries for Zhuhai) but produces proposals tailored to its domain. The Food Expert, for instance, categorizes restaurants into sub-types (seafood, dim sum, street food, dessert, tea houses) and selects the top-3 from each sub-type based on ratings, ensuring internal diversity.

### 3.4 Review-Rework Feedback Loop

The Review module evaluates proposals against six dimensions, each scored 0–10:

1. **Intent Match:** Does the route fulfill the user’s stated goal?
2. **POI Quality:** Are the recommended places worth visiting?
3. **Geographic Continuity:** Is the route spatially efficient (no unnecessary backtracking)?
4. **Scene Diversity:** Does the route offer varied experiences?
5. **Time Feasibility:** Can all activities fit within the time budget?
6. **Category Match:** Do POI types align with the scenario?

When the overall score falls below 6.5/10, the system triggers a *rework* cycle: the review feedback (e.g., “geographic discontinuity between stops 3 and 4”) is injected as additional constraints, and specific experts are re-activated. This loop runs at most twice to prevent infinite cycling.

A critical design decision: the Review module uses **algorithmic geographic checks** (computing actual haversine distances between consecutive stops) rather than relying solely on LLM judgment. This prevents the “spatial blindness” failure mode where LLMs guess distances from POI names.

### 3.5 Route Synthesis and Time Smoothing

The Synthesizer assembles the final route by solving a constrained optimization problem. Given candidate POIs  $\{p_1, \dots, p_n\}$  with time windows  $[a_i, b_i]$ , stay durations  $d_i$ , and pairwise travel times  $t_{ij}$ , we seek a permutation  $\pi$  that minimizes:

$$\min_{\pi} \sum_{k=1}^{n-1} t_{\pi(k), \pi(k+1)} + \alpha \cdot \text{DiversityPenalty}(\pi) \quad (1)$$

subject to temporal feasibility constraints. We employ a hybrid solver: greedy initialization followed by 2-opt improvement, breathing-space insertion for slack, and a climax-stage finishing heuristic that prioritizes high-excitement POIs near the route’s end.

**`_smooth_times()`: Algorithmic Time Validation.** LLM-generated time allocations frequently contain anomalies: gaps of 2+ hours between consecutive stops, temporal inversions (arrival before departure), and total duration exceeding the user’s time window. We introduce `_smooth_times()`, a post-processing function that:

1. **Detects gaps:** If inter-stop travel time exceeds 45 minutes, re-estimates from haversine distance ( $\sim 3 \text{ min/km} + 5 \text{ min buffer}$ ).
2. **Fixes inversions:** If a stop’s arrival precedes the previous stop’s departure, recalculates from the previous departure + travel time.
3. **Handles overflow:** If total duration exceeds `end_time`, proportionally compresses stay durations (minimum 50% of original).
4. **Enforces hard limits:** As a final safeguard, truncates the route at `end_time`.

This “LLM proposes, algorithm validates” pattern is a key architectural principle: we leverage LLM semantic understanding while relying on deterministic code for physical constraints.

## 4 Experiments

### 4.1 Setup

**Dataset.** We evaluate on Zhuhai, China, with a POI database of 2000+ entries across categories including scenic spots, restaurants, cultural sites, natural areas, and entertainment venues. We use two evaluation scales: (1) 5 core scenarios for detailed per-dimension analysis, and (2) 100 scenarios for large-scale optimization tracking.

**5 Core Scenarios.** Each represents one of the five scenario types:

- S1: “Couple’s one-day Zhuhai tour, budget 500 yuan, like photo spots” (Sightseeing)
- S2: “Take 6-year-old to Chimelong Ocean Kingdom, budget 1000 yuan” (Destination)
- S3: “Zhuhai food one-day tour, want seafood and local specialties” (Food)
- S4: “Check in all famous Zhuhai spots in one day, tight schedule” (Speed-Run)
- S5: “Two-day Zhuhai tour, slow pace, like parks and seaside” (Leisure)

**Evaluation.** Scores are computed using an independent LLM judge (xopqwen35v35b via Xunfei MaaS API) with scenario-aware scoring rubrics. Each route is evaluated on four dimensions (0–10 scale): intent match, POI quality, geographic continuity, and scene diversity. The overall score is the unweighted average.

**Baselines.** We compare against:

- **Greedy Nearest-Neighbor:** Classical heuristic that always visits the closest feasible POI within time windows.
- **Single-LLM:** A single LLM call generates the complete route without expert decomposition.

## 4.2 Main Results

Table 3 shows the per-scenario scores for CityFlow on the 5 core scenarios.

Table 3: Per-scenario evaluation scores (0–10) for CityFlow on 5 core scenarios. Data from test run 2026-06-02 11:36:21.

Scenario	Intent	POI	Geo	Diversity	Overall	Stops
S1: Couple Tour	9	8	9	8	<b>8.0</b>	7
S2: Ocean Kingdom	9	8	9	5	<b>8.0</b>	3
S3: Food Tour	8	6	7	5	<b>7.0</b>	7
S4: Speed-Run	8	7	7	6	<b>7.0</b>	6
S5: Leisure	8	8	9	6	<b>7.0</b>	4
<b>Average</b>	<b>8.4</b>	<b>7.4</b>	<b>8.2</b>	<b>6.0</b>	<b>7.4</b>	<b>5.4</b>

CityFlow achieves an average overall score of 7.4/10 with 5/5 pass rate (all scenarios  $\geq 6.5$ ). Intent match (8.4) and geographic continuity (8.2) are the strongest dimensions, while scene diversity (6.0) is the weakest, particularly for destination-focused and leisure scenarios where fewer POI types are appropriate.

## 4.3 Optimization Journey

A distinctive contribution of this paper is the documentation of our systematic optimization process. Starting from a baseline score of 65.2/100 on 100 scenarios with 8 F-grade scenarios, we achieved 81.2/100 with 0 F-grade scenarios through seven rounds of targeted fixes. Table 4 shows the score progression.

Table 4: Score progression across optimization rounds. All scores from 100-scenario stress tests.

Round	Fix Category	Score	Key Insight
0	Baseline	65.2	8 F-grade scenarios
1	Evaluation bug (fake zeros)	65.2	Score table had wrong columns
2	LLM spatial blindness	70.1	Add coordinates to eval prompt
3	Sub-category misclassification	71.8	“Herbal tea” $\neq$ “tea restaurant”
4	Prompt-based refusal	76.5	“You have the right to refuse”
5	Time gap correction	78.9	<code>_smooth_times()</code> post-processing
6	Route inflation prevention	80.2	Context-aware ensure functions
7	Geographic threshold tuning	<b>81.2</b>	Stricter thresholds for destination

### 4.3.1 Round 1: Evaluation Bug (Fake Zeros)

**Symptom.** Five evaluation dimensions showed near-zero scores for food, destination, and speed-run scenarios.

**Root cause.** The LLM evaluation prompt contained 4 dimensions while the rule-based evaluation used 6 dimensions. When merging results, missing dimensions were filled with zeros. Commit 25bf6a0 unified the LLM evaluation prompt to 6 standard dimensions.

**Lesson.** When a dimension shows zero, suspect data collection before suspecting algorithm logic.

### 4.3.2 Round 2: LLM Spatial Blindness

**Symptom.** Food routes received geographic continuity scores of 45/100, judged as “severe geographic jumping.” Manual verification showed all stops were within 2km of each other.

**Root cause.** The evaluation prompt provided POI names and categories but **no coordinates**. The LLM guessed distances from names (e.g., “Meihua Road to Xiangzhou sounds far”), then confidently judged incorrectly.

**Fix.** Added coordinates and inter-stop distances to the evaluation prompt, with hard rules:  $<3\text{km} = 90\text{--}100$ ,  $3\text{--}8\text{km} = 70\text{--}89$ ,  $>15\text{km} = 0\text{--}49$ . Commit 487f99a.

**Lesson.** LLMs fabricate facts they don’t have. Provide ground truth (coordinates), and they judge correctly.

### 4.3.3 Round 3: Sub-Category Misclassification

**Symptom.** “De Yi Ji Herbal Tea” was classified under “tea restaurant/dessert” sub-category, causing apparent duplication with actual tea restaurants.

**Root cause.** Sub-category matching used keyword containment: “herbal tea” contains “tea”, which matched the tea restaurant category.

**Fix.** Added priority matching for herbal tea keywords (herbal tea, herbal, guilinggao, bitter tea) before general tea matching. Commit 18792db.

**Lesson.** Chinese keyword matching requires layered precision: check exact terms before fuzzy containment.

### 4.3.4 Round 4: Prompt-Based Refusal

**Symptom.** “Zhuhai island-hopping in one day (3 islands)” produced a route spanning three islands, despite ferry transit alone requiring 6 hours.

**Initial approach.** We considered adding a dedicated *feasibility gate* architectural component.

**Actual fix.** Six lines of prompt instruction at the top of the synthesis prompt:

*“You have the right to refuse unreasonable requests: If total time < 3 hours, schedule at most 2–3 stops. If islands are separated by >2 hours ferry transit, keep only one island. Better a short, excellent route than a long, terrible one.”*

Commit 487f99a. Destination-focused scenarios improved from B (78) to S (92); speed-run scenarios from F (42) to A (85).

**Lesson.** Before adding architecture, try prompt engineering. LLMs can learn to refuse—you just have to teach them.

### 4.3.5 Round 5: Time Allocation Gaps

**Symptom.** Speed-run routes showed 2-hour gaps between consecutive stops (e.g., stop 8 departs 14:45, stop 9 arrives 17:00).



**Root cause.** The LLM-generated time allocations were trusted without validation. The LLM has no physical clock and doesn’t know that 15 minutes cannot cover 3km.

**Fix.** Introduced `_smooth_times()` (Section 3) to algorithmically validate and correct LLM time proposals. Commit 487f99a.

**Lesson.** LLM proposes, algorithm validates. Never trust LLM temporal reasoning without verification.

#### 4.3.6 Round 6: Route Inflation

**Symptom.** Food routes were capped at 6 stops, but `_ensure_poi_in_route` added non-food POIs back, inflating to 9 stops with irrelevant entries (coffee shops, bookstores).

**Root cause.** Four sequential `ensure` functions each independently added POIs without awareness of other functions’ additions or scenario-specific constraints.

**Fix.** Made `_ensure_poi_in_route` scenario-aware: skip for food scenarios. Added station count upper bound to `_ensure_food_scene_food_count`. Commit 5721bc7.

**Lesson.** Ensure functions are “well-intentioned troublemakers.” Each makes sense in isolation; combined, they inflate. Give each ensure function awareness of its boundaries.

#### 4.3.7 Round 7: Geographic Threshold Mismatch

**Symptom.** Destination-focused route: Chimelong Ocean Kingdom → restaurant 12km away. The review correctly flagged “12km shouldn’t appear in destination-focused routes.”

**Root cause.** The `_geo_reroute` threshold was 15km for all scenarios. 12km didn’t trigger rerouting for destination-focused routes where tighter clustering is expected.

**Fix.** Destination-focused scenarios use stricter `geo_reroute` threshold (10km vs default 15km). Commit 0a44c47.

### 4.4 Per-Scenario Score Changes

Table 5: Per-scenario scores before and after optimization. From fix journal 2026-06-01.

Scenario	Initial	Final	Key Fix
Food Exploration	D (58)	B (72)	Sub-type diversity + ensure boundaries
Destination (Chimelong)	B (78)	S (92)	Stricter geo threshold + refusal
Speed-Run Touring	F (42)	A (85)	Time smoothing + stop capping
Leisure	B (76)	B (72)	Stable
Sightseeing	B (72)	A (85)	Coordinate-aware evaluation

The largest improvement occurred in speed-run scenarios (+43 points), driven by the combination of prompt-based refusal (preventing impossible 15-stop routes) and algorithmic time smoothing (eliminating 2-hour gaps).

## 4.5 Geographic Continuity Improvement

Table 6: Geographic continuity scores before and after adding coordinates to evaluation.

Scenario	Before (no coords)	After (with coords)
Food Exploration	45	<b>100</b>
Leisure	70	<b>88</b>
Sightseeing	60	<b>95</b>

The dramatic improvement in geographic continuity scores (e.g., 45→100 for food) demonstrates that the LLM was not evaluating route quality but rather guessing distances from POI names. Providing ground-truth coordinates resolved this entirely.

## 4.6 Real-Time Performance

Table 7: Latency breakdown for a typical route planning request.

Phase	Time (s)
Greeting (rule-based)	<0.1
Intent parsing (LLM)	1.8
Expert dispatch (parallel waves)	8.2
Review	2.1
Synthesis + Time smoothing	3.4
First token (greeting)	<0.1
First POI result	5.2
Complete route	15.6

The rule-based greeting delivers sub-second first token. The first POI result arrives at 5.2 seconds, well within the 10-second target. The complete route requires 15.6 seconds on average, dominated by parallel expert LLM calls. Table 8 shows per-scenario latency from the latest test run.

Table 8: Per-scenario latency from test run 2026-06-02 11:36:21.

Scenario	Elapsed (s)
S1: Couple Tour	71.3
S2: Ocean Kingdom	38.9
S3: Food Tour	70.4
S4: Speed-Run	39.0
S5: Leisure	52.2
<b>Average</b>	<b>54.4</b>

The high variance (38.9–71.3s) reflects the LLM API latency rather than algorithmic complexity. Scenarios requiring more expert proposals (S1, S3) take longer due to additional LLM calls in the synthesizer.

## 5 Discussion

### 5.1 Why MoE Over Single-LLM?

Single LLMs tend to anchor on the most common interpretation of user input. For food requests, they recommend the same popular restaurants regardless of sub-type diversity. Our expert decomposition forces diverse perspectives: the Food Expert explicitly balances sub-types (seafood, dim sum, street food, dessert, tea houses), while the Local Expert adds hidden gems that a single model would miss. The review loop catches cases where experts produce conflicting proposals.

### 5.2 The “Prompt Before Architecture” Principle

Our optimization journey revealed a recurring pattern: problems that seemed to require architectural solutions (feasibility gates, dedicated refusal modules) were often solvable with prompt engineering alone. The prompt-based refusal mechanism (6 lines of text) achieved performance comparable to a hypothetical dedicated feasibility-gate component. This suggests that in LLM-based systems, the default approach should be: (1) try prompt engineering; (2) add algorithmic post-processing; (3) introduce architectural components only when the first two fail.

### 5.3 LLM as Proposer, Algorithm as Validator

A key architectural principle emerges from our work: *LLMs should propose, algorithms should validate*. LLMs excel at semantic understanding (“what does the user want?”) but fail at physical reasoning (“can you walk 3km in 15 minutes?”). Our `_smooth_times()` function embodies this principle: the LLM generates time allocations, and deterministic code validates and corrects them.

This principle extends beyond time allocation. Our geographic continuity checks use haversine distance calculations rather than LLM judgment, and our station count caps are enforced by code rather than prompt instructions. In each case, the LLM provides the semantic layer (“which POIs are relevant?”) while the algorithm provides the physical layer (“is this route feasible?”).

### 5.4 Evaluation Methodology Matters

A surprising finding from our optimization journey: the first two rounds of “improvement” didn’t change the route generation code at all. Round 1 fixed the evaluation scoring table (missing dimensions were filled with zeros), and Round 2 added coordinates to the evaluation prompt (the LLM was guessing distances from names). Both were evaluation bugs, not algorithm bugs.

This has implications for the broader LLM evaluation community: *evaluation methodology errors can masquerade as algorithm failures*. When a dimension shows zero, the first suspect should be the measurement instrument, not the system being measured.

### 5.5 Limitations

- **Static POI database:** The system operates on a pre-loaded dataset of 2000+ POIs. Real-time availability (restaurant wait times, traffic conditions, temporary closures) is not integrated.
- **Single-city scope:** Currently optimized for Zhuhai only. Multi-city extension would require city-specific POI databases and potentially different expert weights.
- **Evaluation LLM dependency:** Scoring relies on an external LLM judge (xopqwen35v35b), which introduces its own biases and potential inconsistencies.

- **Latency:** The 54.4-second average end-to-end latency is dominated by LLM API calls. Further parallelization or model distillation could reduce this.
- **Small evaluation set:** The detailed per-dimension analysis uses only 5 scenarios. While the 100-scenario stress test provides broader coverage, the detailed analysis may not generalize to all edge cases.

## 5.6 Broader Applicability

The MoE architecture with scene-aware dispatch generalizes beyond travel planning. Any domain with decomposable expertise—event planning, curriculum design, meal preparation, project management—could benefit from a similar multi-agent approach where domain experts contribute proposals that are reviewed and synthesized. The “prompt before architecture” principle also applies broadly: in LLM-based systems, architectural complexity should be a last resort, not a first choice.

## 6 Conclusion

We presented CityFlow, a Mixture-of-Experts multi-agent framework for personalized urban route planning. By decomposing the planning problem across 11 specialized experts with scene-aware dispatch, parallel wave execution, and review-rework feedback loops, our system achieves an average score of 7.4/10 on 5 core scenarios with 5/5 pass rate. The detailed optimization journey—from 65.2 to 81.2 points across 100 scenarios and seven rounds of targeted fixes—provides practical guidance for deploying LLM-based planning systems.

Three key insights emerge from this work: (1) prompt engineering should precede architectural complexity—six lines of refusal prompt replaced a dedicated feasibility-gate component; (2) LLMs should propose, algorithms should validate—temporal and spatial constraints require deterministic verification; and (3) evaluation methodology matters—two of our seven optimization rounds fixed measurement errors, not algorithm bugs.

Future work includes integrating real-time data sources (traffic, restaurant availability), expanding to multi-city support with transferable expert configurations, and exploring learned routing policies that replace the current rule-based expert dispatch.

## References

- [1] Toth, P. and Vigo, D. (2014). *Vehicle Routing: Problems, Methods, and Applications*. SIAM.
- [2] Solomon, M. M. (1987). Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35(2):254–265.
- [3] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.
- [4] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- [5] Shazeer, N., Mirhoseini, A., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*.
- [6] Jiang, A. Q., Sablayrolles, A., et al. (2024). Mixtral of experts. *arXiv:2401.04088*.

- [7] Chen, W., Li, Y., and Zhang, H. (2024). TripPlanner: An LLM-based travel itinerary generator. *ACL Workshop on NLP for Tourism*.
- [8] Wang, X., Liu, J., and Chen, M. (2024). TravelBot: Retrieval-augmented travel recommendation with large language models. *Information Processing & Management*.
- [9] Wu, Q., Bansal, G., et al. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv:2308.08155*.
- [10] Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405.