TripTide: A Benchmark for Adaptive Travel Planning under Disruptions

Anonymous ACL submission

Abstract

Recent work, such as TripCraft and TravelPlanner, has shown the promise of Large Language Models (LLMs) for personalised, constraint-aware travel itinerary generation, but real-world travel often involves disruptions. To address this gap, we introduce TripTide[1], the first benchmark for evaluating LLMs' ability to revise itineraries under realistic disruptions.

TripTide models disruption severity and traveler tolerance, enabling systematic evaluation of LLM responses to events such as transit cancellations, weather closures, and overbooked attractions. We conduct a three-fold evaluation: (i) automatic metrics measuring *Preservation of Intent*, *Responsiveness*, and *Adaptability* (semantic, spatial, and sequential), (ii) an *LLM-as-a-Judge* evaluation, and (iii) a human study assessing revision quality.

Our findings show that LLMs largely preserve semantic and sequential structure, while spatial deviations are higher for shorter itineraries and diminish for longer ones. However, disruption-handling performance declines as itinerary length increases. TripTide[2] provides a foundation for benchmarking robustness and adaptability in LLM-based travel planning.

1 Introduction

Large Language Models (LLMs) have recently been applied to automated travel itinerary generation, with the aim of producing coherent, personalized, and logistically feasible plans utilizing their strong structured reasoning and decision-making abilities. However, despite the increasing use of LLMs for travel itinerary generation, prior work such as *TravelPlanner* (Xie et al., 2024) and *TripCraft* (Chaudhuri et al., 2025) often assumes an ideal, disruption-free environment. Other works,

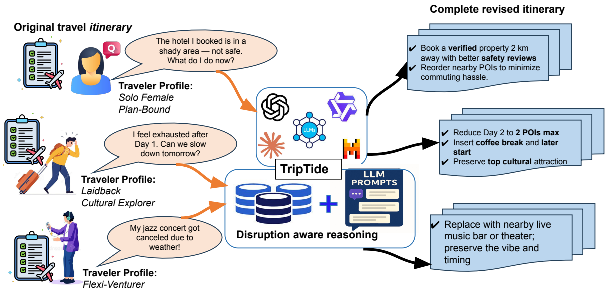

Figure 1: Motivating examples of persona-guided itinerary adaptation in TripTide under different types of disruptions

such as *Flex TravelPlanner* (Oh et al., 2025), have attempted to shed light on the topic but have failed to capture real-life disruption scenarios during real-life travel planning. As a result, they fail to reflect the true complexity of travel planning in the wild. Therefore, it is essential to study how LLMs respond to travel disruptions (such as flight delays, closures of points of interest, accommodation constraints, etc.), offering insights into their real-world utility and ability to revise trips in response to such disruptions.

To fill this critical gap, we augment one of the existing datasets, namely TripCraft (Chaudhuri et al., 2025), which is composed of 1,000 diverse, constraint-rich travel planning queries and introduce TripTide, a new benchmark that simulates a wide range of realistic travel disruptions (see Fig. 1 for motivating examples of TripTide). These disruptions are crafted with carefully designed metadata, including (a) severity levels: step-level (a single activity is impacted), day-level (an entire day's plan is affected), and plan-level (multiple days or the overall itinerary requires revision) and (b) traveler tolerance profiles (Flexi-Venturer or Plan-Bound). Each disruption instance is grounded in plausible scenarios such as cancelled activities, overbooked accommodations, or weather-related closures, creating a high-fidelity environment to evaluate LLMs' capacity for adaptive planning (See Table 1). This

[1] Capturing the ebb and flow of disrupted travel plans
[2] Codebase: https://anonymous.4open.science/r/TripTide-C3A7/

	Subtype	Example / Description
Disruption Category	Transport	Flight Delays, cancellations, missed connections, baggage mishandling, labour strikes, adverse weather conditions, airport closures
	Accommodation	Poor hygiene, unsafe/remote locations, missing promised amenities, overbooking
	Restaurant	Unexpected closures, changes in operating hours, poor hygiene, inadequate dietary support, overcrowding, long waits, surge pricing
	Attractions	Closed for maintenance, weather-related issues, government restrictions, ticket sell-outs, ID rules, overcrowding, limited open days.
	Event/Activity	cancellations, rescheduling, venue changes, ticketing issues, weather or safety risks
	Miscellaneous	traveller illness, jet lag, fatigue, inclement weather, forgotten documents.
Severity	Step-Level	Affects an individual activity or segment
	Day-Level	Impacts all activities on a particular day
	Plan-Level	Necessitates major plan or route overhauls
Traveler Tolerance	Flexi-Venturer	Open to rerouting, substitution, or plan changes
	Plan-Bound	Prefers minimal changes and high plan adherence following Disruption severity.

Table 1: Taxonomy of disruption categories, severity levels, and traveler tolerance profiles

augmented dataset enables us to systematically investigate LLMs' performance in scenarios that require quick thinking and plan revision, with a deep understanding of user preferences.

To assess how effectively LLMs handle travel disruptions, we propose a suite of three novel evaluation metrics tailored to this dynamic context. The 'Preservation of Intent' score assesses the extent to which the modified itinerary upholds the essential objectives and experiential priorities of the original plan. The 'Responsiveness' score measures the model's ability to promptly and appropriately respond to the disruption, distinguishing between decisive interventions and vague or evasive modifications. In addition, we introduce a suite of 'Adaptability' metrics that quantify the semantic, spatial, and sequential shifts between the original and revised itineraries, thereby offering detailed insights into how the model restructures plans under evolving constraints. We also employ the LLM-as-a-Judge method for automatic evaluation of the generated plans and subsequently validate the experimental results by correlating them with manual evaluations of the disruption-mitigated plans conducted by human experts. We found that language models effectively acknowledge disruptions and react to mitigate them; however, they sometimes struggle to maintain narrative coherence between initial and revised plans.

Our work makes the following key contributions: (1) **LLM-Based Disruption Response Generation:** We introduce TripTide, the first benchmark specifically designed to evaluate the robustness of LLM in travel planning under realistic disruptions. We develop a few-shot prompt-driven mechanism to handle different types of disruption by guiding LLMs to generate coherent, minimally altered revised itineraries. (2) **Structured Annotation and Comprehensive Evaluation Metrics:** We contribute structured, annotated, revised plans, curated post-disruption, which serve as a benchmark for evaluating travel planning systems. Additionally, we also propose a set of novel evaluation metrics that measure the preservation of user intent, responsiveness to disruptions and adaptability of the revised plans. (3) **Analysis of LLM Planning Behavior under Disruption:** We identify key behavioral patterns in how LLMs handle disruptions, providing insights into their underlying planning heuristics and standard failure modes.

2 Related Work

LLMs in Planning tasks.

With the advent of LLMs, there has been a growing interest in their ability to perform such planning through natural language (Valmeekam et al., 2023; Bohnet et al., 2024). LLMs have shown impressive capabilities in task decomposition, commonsense reasoning, and step-by-step decision-making (Wei et al., 2022; Yang and Tomar, 2023), making them compelling candidates for real-world planning applications. Recent approaches such as *LLM-Planner* (Song et al., 2023) enable few-shot grounded planning by iteratively refining high-level plans in response to feedback. At the same time, hybrid methods integrating LLMs with classical algorithms like Monte Carlo Tree Search have demonstrated improved planning efficiency (Zhao et al., 2024; Świechowski et al., 2023). Other existing works like *TripTailor* (Shen et al., 2025) and (Shao et al., 2025) extending the *TravelPlanner* idea talk about the LLM's reasoning capabilities across the China dataset. Despite this progress, LLMs still struggle to reliably generate robust and coherent plans in open-ended environments (Zhang et al., 2025), often faltering in sub-goal coordination and long-horizon dependencies (Valmeekam et al., 2023; Kambhampati et al., 2023; Singh et al., 2024). While chain-of-thought prompting (Wei et al., 2022) and fine-tuning improve benchmark performance, they falter on novel, complex tasks. Retrieval-based methods such as RAG (Ni et al., 2025) and knowledge graphs (Song et al., 2024; Xiao et al., 2020) likewise remain unproven in real-world travel disruption scenarios.

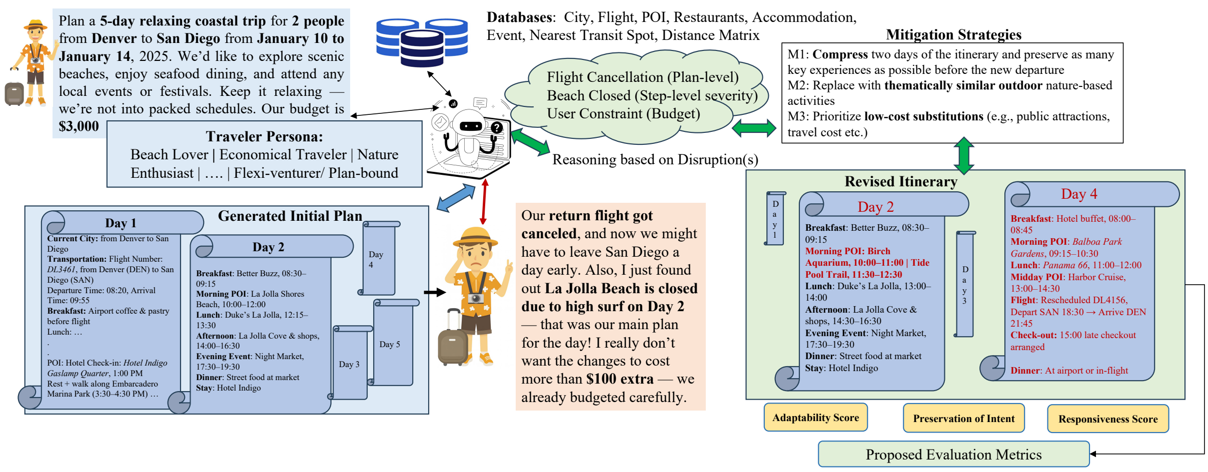

Figure 2: Illustration of multi-level disruption handling and adaptive travel replanning by TripTide

Travel Planning with LLMs. Travel planning poses a multifaceted challenge, involving the coordination of diverse sub-goals such as time management, cost optimization, and personalized user preferences (Gurjar and Gupta, 2021). The recent application of LLMs to this domain has sparked interest due to their ability to interpret and generate natural language instructions (Xi et al., 2025; Jonnala et al., 2025). Initial benchmarks, such as the one proposed by Xie et al. (2024), highlight the difficulty LLMs face in balancing multiple constraints (Borro et al., 2025), often producing itineraries that violate feasibility or preference criteria. Although subsequent works (Hao et al., 2025; Gundawar et al., 2024) report promising results, they typically oversimplify the problem by omitting real-world limitations such as transit schedules, event timing, and dynamic user contexts (Shao et al., 2024) and using discrete binary evaluation metrics. Although recent efforts (Chaudhuri et al., 2025; Chen et al., 2024; Singh et al., 2024) have advanced the use of continuous metrics to evaluate rationality, personalization, and alignment of user preference in travel plans, they largely overlook the disruption-prone nature of real-world travel.

Disruptions in Travel Planning. Despite travel being inherently susceptible to disruptions such as weather events, transportation delays, and overbooked attractions, prior studies have largely ignored this aspect of planning. Existing benchmarks and systems assume ideal execution of generated itineraries, without considering how plans should evolve when unexpected obstacles arise. To the best of our knowledge, TripTide is the first benchmark to systematically study how LLMs respond to travel disruptions. TripTide simulates a diverse range of real-world disruptions and introduces metrics to assess how well LLMs adapt while preserving user goals and contextual feasibility.

3 TripTide Dataset Curation

TripTide consists of 1,000 travel planning queries (adapted from TripCraft) across three trip durations (3-day, 5-day, and 7-day), each paired with a corresponding disruption query and a human-annotated revised plan. These revised plans are designed to respect both the traveler's persona and disruption tolerance level (e.g., 'Flexi-Venturer' or 'Plan-Bound'), ensuring that LLMs are evaluated for their ability to maintain personalization and coherence under restrictions. Fig. 2 depicts the block diagram of TripTide illustrating multi-level disruption(s) handling and replanning features of our proposed system.

3.1 Disruption Generation

Task Overview: For each of the annotated plans provided in the TripCraft dataset (Chaudhuri et al., 2025), we generate three potential disruptions per day of the itinerary by prompting GPT-4o and Gemini 2.5 Pro (Comanici et al., 2025). Each disruption is then paired with a corresponding disruption query, which is used as input to prompt GPT-4o[3] to generate a persona-aware revised plan. The complete prompts are mentioned in our codebase[2]. Hyperparameter details are provided in Appendix B.

Disruption Categories and Scenarios. Table 1 outlines the various categories of disruptions that may occur during a trip, along with representative scenarios under each category.

Disruption Severity Levels. As a key feature of

[3] https://openai.com/index/gpt-4o-system-card/

our benchmark, we categorize disruptions based on their scope into three levels: *Step-Level*, *Day-Level*, and *Plan-Level*. **Step-Level** disruptions affect a single activity or step within the itinerary (e.g., a visit to a specific attraction or a meal reservation). **Day-Level** disruptions impact the set of activities scheduled for a particular day, requiring partial rescheduling or substitution of that day's Point-of-Interest (POI) list. **Plan-Level** disruptions have broader consequences, potentially requiring significant changes across multiple days of the itinerary, including complete reordering or replacement of the POI list.

User Personas and Tolerance Levels. We categorize travelers using structured user personas that encapsulate their goals, preferences, and behavioral traits during a trip. Examples of such personas include the *Adventure Seeker*, the *Cultural Explorer*, the *Economical Traveler*, or the *Mountain Enthusiast*, each reflecting distinct travel styles and motivations. Each persona in our framework is further associated with a predefined tolerance level to disruption, ranging from highly flexible to strictly structured, based on factors such as adaptability, planning preferences, and sensitivity to change.

We introduce a novel, first-of-its-kind modeling of disruption tolerance in travel planning, distinguishing between two representative user types: (1) *Flexi-Venturer*, who readily adapts to changes and show low sensitivity to deviations263 from the original plan, and (2) *Plan-Bound* traveler, who is rigid and prefers minimal deviation from the original itinerary. While the user persona reflects the traveler's preferences, such as whether they seek relaxation or adventures, the disruption tolerance level indicates how well the traveler adapts to disruptions. Incorporating the tolerance level alongside the user persona enables a more accurate representation of travelers' real-world behaviors under disruptions.

3.2 Annotations and Quality Control

Human-Level Annotation and Disruption Query Generation. To ensure realistic modeling of disruptions in travel itineraries, we adopt a rigorous human-in-the-loop annotation framework. For each day of travel, we identify three plausible disruptions, resulting in a comprehensive pool of disruption scenarios per plan. Specifically, for the 344 three-day plans, this yields 2489 disruption cases. Similarly, disruptions are generated for all five-day and seven-day plans using the same per-day logic. Table 2 contains the detailed distribution per disruption category. In total, our dataset includes **11,058** possible disruptions spanning a diverse range of conditions, traveler profiles, and destinations.

Human annotators[4] were then tasked with sampling the most contextually meaningful disruption for each plan. They carefully reviewed the full reference itinerary and selected disruptions that emphasized realism, diversity (in terms of type and severity), and traveler-specific relevance. For each selected disruption, annotators collaborated with the GPT-4o and Gemini 2.5 Pro models to generate natural language queries that emulate how travelers would express real-world concerns (e.g., *"My flight's been delayed. What now?"*). Annotators then revised the original itinerary in response to the disruption, ensuring logical adaptation, temporal feasibility, and overall trip coherence. That helped to encode grounded, expressive, and context-rich examples that reflect how disruptions are perceived and handled in real-world travel. Detailed annotation guidelines are in Appendix C.

Constraint Dimension	Subtype	Plan Duration			Total
		3-day	5-day	7-day	
Disruption Category	Transport	118	364	555	1,037
	Accomodation	225	416	418	1,059
	Restaurant	1,072	1,223	1,618	3,913
	Attractions	980	1,223	1,618	3,821
	Miscellaneous	94	437	697	1,228
	Total	**2,489**	**3,663**	**4,906**	**11,058**
Disruption Severity	Step-level	111	140	134	385
	Day-level	117	111	131	359
	Plan-level	116	73	67	256
	Total	**344**	**324**	**332**	**1,000**
Traveler Tolerance	Flexi-Venturer	172	162	166	500
	Plan-Bound	172	162	166	500
	Total	**344**	**324**	**332**	**1,000**

Table 2: Distribution of disruptions by category, severity, and traveler tolerance across different plan durations.

Automated Script-Based Verification.

To complement human annotations with structural and semantic correctness, we implemented a robust script-based validation pipeline. The primary goal was to ensure that all entities mentioned in the revised plan, such as accommodations, flights, events, and attractions, were strictly derived from the original reference itinerary. The script conducts thorough entity alignment checks and flags any hallucinated or mismatched elements that are not grounded in the source data.

Additionally, the script ensures logical consistency across time, location, and activity sequences. It verifies that there are no temporal conflicts (e.g.,

[4]Undergraduate interns at our lab

double-booked time slots), that inter-city transitions are feasible, and that each revised plan respects the budget and structural constraints of the original. This automated check serves as a second layer of quality assurance, substantially reducing annotation noise. Together, our three-tiered curation strategy, combining expert human annotation with rule-based verification, followed by LLM-as-a-Judge evaluation alongside assessments from human experts, ensures a high-integrity dataset for benchmarking disruption-aware travel planning systems.

Disruption Categorization and Tolerance Profiling. To model user preferences and plan flexibility, each disruption is categorized into a level, whereas the traveler is tagged with a tolerance level, reflecting how tolerant a traveler might be toward changes. Table 2 shows the distribution of tolerance levels and disruption severity across different plan durations.

4 Automated Evaluation

4.1 Evaluation Metrics

Given inputs (current plan, disruption, its severity, and tolerance), the model generates an alternative feasible plan. We propose evaluation of this new plan using three novel metrics: *preservation of intent*, *responsiveness*, and *adaptability*.

1. Preservation of Intent. This metric evaluates whether the user's original travel intent, as characterized by constraints and preferences, is retained in the revised plan. We apply the commonsense and hard constraint pass rates (CPR and HCPR), delivery rate, and final pass rate introduced by TravelPlanner (Xie et al., 2024) to the final alternative plan to estimate the percentage of intent preserved. Higher values are preferred.

2. Responsiveness. This measures whether the model acknowledges the disruption and makes meaningful alterations to the original plan. For each day, the responsiveness rate is given by the ratio of fraction of total plans that were mitigated. Higher values are preferred.

3. Adaptation Quality. This measures the quality of adaptation, specifically, whether the revised plan offers a realistic and contextually appropriate substitute or reorganization. It evaluates if the model suggests thematically consistent alternatives while minimizing disruptions. Each score is finalized by taking the absolute difference between the initial plan and the revised plan score. Thus, a lower value of the metric score reflects that the revised plan is more thematically consistent with minimal changes. The score is decomposed into three components: semantic, spatial, and sequential. Lower values are preferred for all three.

i) Semantic Adaptability (A_{sem}). This component measures how thematically consistent the revised plan is with the original plan. Specifically, it quantifies the semantic closeness between PoIs in the initial and revised itineraries with the user persona using BERT-based cosine similarity. While we adopt the persona score formulation introduced in (Chaudhuri et al., 2025), in our setting, it is repurposed to reflect plan-level thematic continuity rather than alignment with the user persona:

$$\bar{S}_{\text{ps}} = \frac{1}{M \cdot N} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\mathbf{p}_j \cdot \mathbf{q}_i}{\|\mathbf{p}_j\| \|\mathbf{q}_i\|} \tag{1}$$

where $\mathbf{p}_j$ and $\mathbf{q}_i$ are BERT embeddings of the j-th persona component and i-th PoI name respectively, M is the number of persona components, and N is the number of PoIs in the plan. The semantic adaptability is computed as $A_{\text{sem}} = |\bar{S}^{\text{i}}_{\text{ps}} - \bar{S}^{\text{r}}_{\text{ps}}|$ where $\bar{S}^{\text{i}}_{\text{ps}}$ and $\bar{S}^{\text{r}}_{\text{ps}}$ are the persona scores of the initial and revised plans, respectively.

ii) Spatial Adaptability (A_{spa}). This measures the difference in spatial convenience between the original and revised plans. Following (Chaudhuri et al., 2025), we compute the spatial score for each plan using:

$$S_s(d) = \begin{cases} 1 - 0.5\left(\frac{d}{d_0}\right), & \text{if } d \le d_0 \\ 0.5 \exp\left(-\lambda(d - d_0)\right), & \text{if } d > d_0 \end{cases} \tag{2}$$

$$\bar{S}_{\text{spatial}} = \frac{1}{N} \sum_{i=1}^{N} S_s(d_i) \tag{3}$$

where d_i is the distance of the i-th PoI from the nearest public transit station, d_0 is the distance threshold, and λ is the decay rate. Therefore, spatial adaptability is computed as $A_{\text{spa}} = |\bar{S}^{\text{i}}_{\text{spatial}} - \bar{S}^{\text{r}}_{\text{spatial}}|$ where $\bar{S}^{\text{i}}_{\text{spatial}}$ and $\bar{S}^{\text{r}}_{\text{spatial}}$ are the spatial scores of the initial and revised plans, respectively.

iii) Sequential Adaptability (A_{seq}). This evaluates changes in the order of PoIs across days using the normalized edit distance:

$$A_{\text{seq}} = \frac{\text{ED}(\mathcal{G}, \mathcal{A})}{\max(|\mathcal{G}|, |\mathcal{A}|)} \tag{4}$$

where $\mathcal{G}$ and $\mathcal{A}$ are PoI sequences of the initial and revised plans, and $\text{ED}(\cdot, \cdot)$ is the edit distance. We

Model Name	Plan Duration	Delivery Rate (%)↑	CPR (%)↑		HCPR (%)↑		Final Pass Rate (%)↑	Adaptability (%)			Responsiveness Rate (%)↑
			Micro	Macro	Micro	Macro		Semantic (A_{sem})↓	Spatial (A_{spa})↓	Sequential (A_{seq})↓	
GPT-4o	3-day	100.00	91.02	29.95	50.93	48.84	29.37	0.34	1.22	10.89	89.53
	5-day	99.69	**91.46**	**43.21**	**67.38**	**67.59**	**41.98**	**0.02**	0.66	**1.86**	81.79
	7-day	100.00	87.35	33.14	55.59	57.54	32.54	0.20	**0.30**	5.58	79.82
Qwen2.5-7B-Instruct	3-day	100.00	89.56	31.16	54.02	51.93	29.98	0.67	5.94	56.26	97.67
	5-day	100.00	89.05	29.42	58.77	59.76	29.12	0.04	1.25	28.46	82.02
	7-day	97.54	80.98	16.26	25.16	29.15	16.27	1.24	2.10	24.04	88.67
Phi-4 mini Instruct	3-day	45.23	39.75	18.68	20.36	24.49	18.26	26.27	3.32	26.64	75.94
	5-day	83.14	68.81	6.18	18.89	13.28	5.87	7.01	1.45	30.13	50.31
	7-day	100.00	73.43	3.06	7.67	7.96	3.06	30.34	27.78	42.74	**100.00**

Table 3: Metric scores for preservation of user intent (Delivery, CPR, HCPR, Final Pass), Adaptability (A_{sem}, A_{spa}, A_{seq}), and Responsiveness across models and plan durations. Differences are absolute and scaled to % points.

average this over all days of both the plans to get A_{seq}.

4.2 Results

We report the performance[5] of GPT-4o and Qwen2.5-7B-Instruct and Phi-4-mini Instruct across initial and revised itineraries using *Preservation of Intent* metrics, *Adaptability* metrics (*semantic*, *spatial*, *sequential*), and *Responsiveness* in Table 3 across three plan durations: 3-day, 5-day, and 7-day trips.

Table 3 shows clear model trade-offs across plan durations. GPT-4o is the most stable overall: it achieves near-perfect delivery rates, the highest final pass rates (peaking at 41.98% for 5-day), and consistently low semantic, spatial, and sequential drift, indicating strong preservation of user intent under disruptions. Performance degrades mildly with longer horizons, reflected in declining responsiveness.

Qwen2.5-7B-Instruct matches GPT-4o on delivery rate and semantic adaptability for shorter plans and shows strong responsiveness, especially for 7-day itineraries. However, its constraint satisfaction (HCPR, final pass) degrade notably with plan length, driven by aggressive re-optimization. Its sequential adaptability is also significantly worse than GPT-4o.

Phi-4-mini-Instruct exhibits unstable behavior: delivery and responsiveness rates improve with longer plans (reaching 100% for 7-day), but CPR/HCPR, final pass rates, and adaptability (especially semantic and sequential) collapse, indicating poor intent and constraint preservation despite detecting disruptions.

Overall, GPT-4o offers the best balance of intent preservation and adaptability, Qwen2.5-7B scales better in responsiveness at the cost of structural stability, and Phi-4-mini struggles with reliable constraint reasoning. We present detailed analysis of these results in Appendix D. We also present disruption severity based analysis in Appendix E.

5 LLM-as-a-Judge based Evaluation

Following the *LLM-as-a-Judge* paradigm (Zheng et al., 2023), beyond the automated metrics in Section 4, we further assess revised plans generated by GPT-4o using an independent LLM. We choose GPT-4o as it achieves the best average performance across metrics (Table 3). For each query, the judge is provided with the original plan, disruption details, and the revised plan. The revised itinerary is rated on a 1–5 scale (1 for minimal or ineffective revision; 5 for complete resolution of the disruption). Table 8 in Appendix F shows scoring rubric. We conduct this evaluation using Llama-3.1-8B-Instruct[6] with a carefully designed prompt (See Appendix A.5) to reduce evaluation bias and hallucination.

Fig. 3 shows a clear pattern: 3-day plans are judged strongest overall (highest mean; majority *Good* with a noticeable number of *Excellent* samples). 5-day skews toward *Average*/*Good* with very few *Excellent*, yielding the lowest mean. 7-day sits between: more *Good* than 5-day, but still limited *Excellent*.

Relation to automated metrics. Comparing results in Fig. 3 with the automated metric scores in Table 3 , we see agreement on overall plan quality (few outright failures). Delivery rate is saturated (≈100%), so differences emerge in constraint adherence. 5-day plans, despite having the lowest judge mean, achieve the strongest compliance: HCPR (micro/macro) 67.38/67.59 and Final-Pass

[5]We evaluate the models on the entire dataset due to the absence of a predefined train-test split. These evaluations are diagnostic in nature and are intended to inform design insights rather than generalization performance.

[6]https://huggingface.co/meta-llama/Llama-3.1-8B

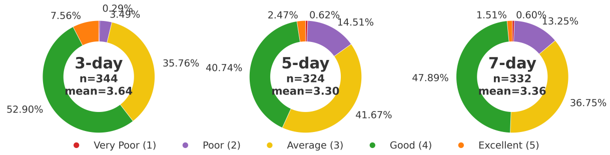

Figure 3: Llama-3.1 8B Instruct Evaluation Results

Rate 41.98%. 3-day plans, while best judged (mean 3.64), show weaker formal validity: $\text{CPR}_{\text{macro}}$ of 29.95, $\text{HCPR}_{\text{macro}}$ of 48.84, and Final-Pass Rate of 29.37%. The 7-day bucket is intermediate: HCPR micro/macro of 55.59/57.54 and Final-Pass Rate of 32.54%. Taken together, the 'Llama-as-a-Judge' ratings reflect perceived quality (favoring 3-day), while automated scripts surface hard feasibility (favoring 5-day), with 7-day providing a balanced middle ground.

6 Human Evaluation of Revised Plans

We complement automated evaluation with a human evaluation of disruption-aware itinerary revisions. Three domain experts independently reviewed 100 revised itineraries, sampled across 3-, 5-, and 7-day plans. Reviewers examined the original plan, disruption details, and revised itinerary, assessing responsiveness and preservation of semantic, spatial, and sequential coherence. In this section, we analyze how GPT-4o handled unexpected disruptions in terms of reasoning, contextual understanding, and adaptability. Representative case studies are in Appendix G.

6.1 General trends

Reliable disruption detection, disciplined yet non-trivial revisions. GPT-4o planner consistently identified the disrupted element and initiated a corrective edit. While many fixes were intentionally local (high responsiveness), a measurable subset propagated beyond the impacted slot or day, indicating a deliberate but imperfect balance between targeted repair and broader itinerary stability.

Quality changes with trip length. In 3-day itineraries, edits remained compact with limited downstream consequences, enabling strong sequentiality and semantic preservation. 5-day plans more often incurred scope drift (additional edits not strictly required by the disruption), reflecting the added complexity of multi-day dependencies. 7-day itineraries showed the widest variance: several exemplary, end-to-end consistent repairs, but also more opportunities for spatial or temporal slippage when local changes were not fully propagated.

6.2 Where the Planner did well

Smart swaps. When a POI was closed, mispriced, or poorly located, the system selected a like-for-like alternative (similar category, nearby geography, compatible opening hours). This preserved semantic fidelity (intent and activity type), maintained spatial coherence (short transfers, clustered movement), and limited ripple effects, supporting sequentiality by keeping rest of the day intact.

Human factors. Experts observed fatigue-aware behavior: after long drives or dense activity blocks, the planner often inserted lighter follow-ups or buffer time. Such adjustments improved sequentiality (realistic pacing) and responsiveness (addressing the root strain) without diluting the day's semantic goals.

Logistics. The system closed execution gaps (e.g., adding appropriate-capacity ground transport for larger parties or linking airport legs to city transfers), ensuring end-to-end feasibility. These edits improved spatial and sequential coherence while remaining minimally invasive

Persona fit. Substitutions typically reflected stated interests and constraints (e.g., nature-aligned alternatives for "Nature" travelers; shopping added where requested), sustaining semantic alignment with preferences while respecting spatial and temporal realities.

6.3 Where the Planner struggled

Superficial fixes. A subset of cases exhibited nominal edits (e.g., renaming lodging or swapping to a venue with similar limitations) that did not materially resolve the underlying constraint. Such changes preserved surface semantics but weakened responsiveness and yielded little gain in spatial or sequential feasibility.

Missing the root cause. After very long transfers, the planner sometimes addressed a symptom (e.g., pushing a meal) without rebalancing the next day to account for accumulated fatigue. This under-correction harmed sequentiality (unrealistic pacing) and occasionally the day's semantic goals, despite appearing minimal on paper.

Real-world timing. Sometimes, replacements overlooked operating-hour constraints, substituting a closed venue with another that was also unavailable at the proposed time. These errors broke sequentiality and spatial realism, despite intent to preserve activity type.

Ripple effects. Local adjustments (e.g., "leave earlier", "add generic rest stops") were not always propagated to dependent activities, leaving downstream segments compressed or misordered. Such partial propagation reduced sequentiality and sometimes undermined responsiveness, even when the initial substitution appeared reasonable.

6.4 Analysis across Disruption Types

Accommodation Disruptions. In most cases (4/5), unsuitable accommodations were replaced with better alternatives aligned with traveler preferences (e.g., size, comfort or luxury). In one instance, the itinerary was adjusted instead to account for fatigue from a late check-in. These responses demonstrate strong responsiveness, though minor drops in sequentiality occurred when overnight changes altered the following day's schedule.

Transportation Disruptions. Transportation issues were resolved with limited impact on itinerary structure by adjusting departure times, adding rest breaks, or switching transport modes (e.g., public to self-driven). These changes preserved feasibility, semantic alignment, and spatial coherence, showcasing a strong contextual understanding, though occasional conservative decisions led to slightly lower responsiveness for minor disruptions.

Attraction Disruptions. Mitigation strategies for disrupted attractions followed a two-fold approach: replacing the activity or rescheduling it to a more suitable time. For example, closed or contextually inappropriate attractions (e.g., hiking in low light) were either swapped for alternatives or moved to earlier in the day. These adaptations maintained both user's intent and temporal structure, yielding high semantic and sequentiality scores. Rarely, spatial score declined marginally when replacements introduced longer travel distances.

Miscellaneous Disruptions. A broad range of mitigation responses were observed for disruptions related to restaurants, fatigue, and profile mismatches. For example, closed dining venues were substituted with similar alternatives, while travelers experiencing fatigue were assigned more relaxed schedules, including later starts and reduced physical demands. Profile-aware adjustments (e.g., budget vs. luxury) improved semantic alignment, though spatial efficiency occasionally decreased when suitable alternatives were farther away.

Analysis of Examples. The two snippets in Figs. 4 and 5 show the robust performance of GPT-4o for a 5-day and a 3-day case from our dataset. The model makes multiple semantic mistakes in 5-day

Traveler Persona: {"type": "Laidback", "purpose": "Nature","cuisine":"Mexican, Italian", "budget": "Economical", "pref": "Beaches","Tolerance": "Flexi-Venturer"}
Original Plan: {"day": 1, ... } {"day": 2, ... } {"day": 3, ... } {"day": 4, "city": "Durango", "transport": "-",, "breakfast": "-", "attraction": "-", "lunch": "Zia Taqueria, Durango", "dinner": "-", "stay": "Modern Victorian"}
Disruption: {"Day 4 Lunch": "Zia Taqueria, Durango", "Disruption Category':" Restaurants ", "Reason":"The Zia taquerias Tuesday deals attract college students, leading to long waits and limited seating for groups."}
Revised Plan: {"day": 1, ... } {"day": 2, ... } {"day": 3, ... } {"day": 4, "city": "Durango", "transport": –, "breakfast": "Durango Doughworks", "attraction": "Garden of the Gods", "lunch": "Michel's Corner Crepes, Durango", "dinner": "Mountain Shadows", "stay": "Modern Victorian"}
Semantic Issue: The Restaurant Disruption in Day 4 is handled by the LLM generated plan, but it fails to capture the cuisine preference of the traveler that is Mexican or Italian, rather it replaces a Mexican style restaurant by a French type restaurant.

Figure 4: Semantic issues for a 5-day plan

Traveler Persona: {"type": "Laidback", "purpose": "History", "budget": "Economical", "pref": "Museums","Tolerance": "Plan-Bound"}
Initial Plan: {"day": 1, "city": "San Diago", "poi_sequence": ["Calibro National Monument", "Coronado Bridge' ...]}
{"day": 2, "city": "San Diago", "poi_sequence": ["Seaport Village (377.43m away)", "Petco Park" ...]}
{"day": 3, "city": "San Diago", 'poi_sequence": ["Return to Seattle by flight number F0293908,[11:35, 13:05]"]}
Disruption: "Day 2 Attraction": "Seaport Village", "Disruption Category":"Attraction", "Reason": "There is an ongoing plan called Seaport San Diego to redevelop Seaport Village, because of which tourists are not allowed."
Revised Plan: {"day": 1, "city": "San Diago", "poi_sequence": ["Calibro National Monument", "Coronado Bridge' ...]}
{"day": 2, "city": "San Diago", "poi_sequence": ["La Jolla Shores Park (243.08m away)", "Petco Park" ...]}
{"day": 3, "city": "San Diago", 'poi_sequence": ["Return to Seattle by flight number F0244163 [12:37, 13:52]"]}
Spatial Adaptability Even with the attraction swap and flight changes the distances of POIs to nearest transit stayed within similar ranges (50-200m), thus preserving the spatial adaptability A_{spa}.

Figure 5: Spatial Preservation for a 3 day plan

while maintaining the spatial coherence for 3-day.

7 Conclusion

We introduce TripTide, the benchmark to evaluate LLMs' ability to prescribe travel itineraries, in the face of real-world disruptions. To the best of our knowledge, this is the first work to integrate disruption severity levels with user-specific tolerance profiles in the context of travel planning. By jointly modeling the granularity of disruptions (*step-level*, *day-level*, and *plan-level*), alongside traveler flexibility types (e.g., *Flexi-Venturer* vs. *Plan-Bound*), our framework captures a significantly broader and more realistic spectrum of traveler behaviors. Moreover, alongside LLM-act-as-a Judge and manual evaluation by human experts, we introduce a set of novel evaluation metrics to evaluate LLM-generated plans, via measuring preservation of user intent, the responsiveness to disruptions, and the adaptability of the revised plans. This nuanced integration facilitates personalized and context-aware itinerary adaptations, effectively bridging a critical gap in prior work, which does not care for disruptions.

8 Limitations

While TripTide provides a structured benchmark for evaluating disruption-aware itinerary revision, several limitations remain. The disruptions are synthetically constructed to enable controlled analysis, but may not fully capture the cascading or evolving nature of real-world travel disruptions. In addition, the benchmark does not incorporate real-time external data (e.g., live flight status or weather updates), and thus emphasizes reasoning-based adaptation over factual verification. Finally, itinerary revisions are evaluated as single-step responses, whereas real-world travel planning often involves multi-turn interactions. Extending TripTide to support interactive revision workflows and real-world feedback remains an important direction for future work.

9 Ethical Considerations

This work benchmarks LLMs for disruption-aware travel itinerary revision using fully synthetic scenarios and does not involve personal or sensitive user data. While intended solely for research evaluation, real-world deployment of such systems may influence user decisions during disruptions; therefore, model outputs should be treated as decision support and used with appropriate user oversight.

References

Bernd Bohnet, Azade Nova, Aaron T Parisi, Kevin Swersky, Katayoon Goshvadi, Hanjun Dai, Dale Schuurmans, Noah Fiedel, and Hanie Sedghi. 2024. Exploring and benchmarking the planning capabilities of large language models. *Preprint*, arXiv:2406.13094.

Andrey Borro, Patricia J Riddle, Michael W Barley, and Michael J Witbrock. 2025. Large language models as common-sense heuristics. *Preprint*, arXiv:2501.18816.

Soumyabrata Chaudhuri, Pranav Purkar, Ritwik Raghav, Shubhojit Mallick, Manish Gupta, Abhik Jana, and Shreya Ghosh. 2025. Tripcraft: A benchmark for spatio-temporally fine grained travel planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17035–17064, Vienna, Austria. Association for Computational Linguistics.

Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. 2024. Travelagent: An ai assistant for personalized travel planning. *Preprint*, arXiv:2409.08069.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Atharva Gundawar, Mudit Verma, Lin Guan, Karthik Valmeekam, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Robust planning with llm-modulo framework: Case study in travel planning. *arXiv preprint arXiv:2405.20625*.

Omkar Gurjar and Manish Gupta. 2021. Should i visit this place? inclusion and exclusion phrase mining from reviews. In *European Conference on Information Retrieval*, pages 287–294. Springer.

Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. 2025. Large language models can solve real-world planning rigorously with formal verification tools. *Preprint*, arXiv:2404.11891.

Ramya Jonnala, Gongbo Liang, Jeong Yang, and Izzat Alsmadi. 2025. Exploring the potential of large language models in public transportation: San antonio case study. *arXiv preprint arXiv:2501.03904*.

Subbarao. Kambhampati, Karthik. Valmeekam, Matthew. Marquez, and Lin. Guan. 2023. On the role of large language models in planning. Tutorial presented at the International Conference on Automated Planning and Scheduling (ICAPS), Prague.

Hang Ni, Fan Liu, Xinyu Ma, Lixin Su, Shuaiqiang Wang, Dawei Yin, Hui Xiong, and Hao Liu. 2025. Tp-rag: Benchmarking retrieval-augmented large language model agents for spatiotemporal-aware travel planning. *Preprint*, arXiv:2504.08694.

Juhyun Oh, Eunsu Kim, and Alice Oh. 2025. FLEX-TRAVELPLANNER: A BENCHMARK FOR FLEXIBLE PLANNING WITH LANGUAGE AGENTS. In *Workshop on Reasoning and Planning for Large Language Models*.

Jie-Jing Shao, Xiao-Wen Yang, Bo-Wen Zhang, Baizhi Chen, Wen-Da Wei, Lan-Zhe Guo, and Yu-feng Li. 2024. Chinatravel: A real-world benchmark for language agents in chinese travel planning. *arXiv preprint arXiv:2412.13682*.

Jie-Jing Shao, Bo-Wen Zhang, Xiao-Wen Yang, Baizhi Chen, Si-Yu Han, Wen-Da Wei, Guohao Cai, Zhenhua Dong, Lan-Zhe Guo, and Yu-Feng Li. 2025. Chinatravel: An open-ended benchmark for language agents in chinese travel planning. *Preprint*, arXiv:2412.13682.

Yuanzhe Shen, Kaimin Wang, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2025. Triptailor: A real-world benchmark for personalized travel planning. *Preprint*, arXiv:2508.01432.

Harmanpreet Singh, Nikhil Verma, Yixiao Wang, Manasa Bharadwaj, Homa Fashandi, Kevin Ferreira, and Chul Lee. 2024. Personal large language model agents: A case study on tailored travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 486–514, Miami, Florida, US. Association for Computational Linguistics.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Sihan Song, Chuncheng Yang, Li Xu, Haibin Shang, Zhuo Li, and Yinghui Chang. 2024. Travelrag: A tourist attraction retrieval framework based on multi-layer knowledge graph. *ISPRS International Journal of Geo-Information*, 13(11).

Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. 2023. Monte carlo tree search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56(3):2497–2562.

Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. 2023. On the planning abilities of large language models (a critical investigation with a proposed benchmark). *Preprint*, arXiv:2302.06706.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Dinghe Xiao, Nannan Wang, Jiangang Yu, Chunhong Zhang, and Jiaqi Wu. 2020. A practice of tourism knowledge graph construction based on heterogeneous information. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 939–949, Haikou, China. Chinese Information Processing Society of China.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. In *Forty-first International Conference on Machine Learning*.

Yunhao Yang and Anshul Tomar. 2023. On the planning, search, and memorization capabilities of large language models. In *International Conference on Intelligent Vision and Computing*, pages 24–38. Springer.

Cong Zhang, Xin Deik Goh, Dexun Li, Hao Zhang, and Yong Liu. 2025. Planning with multi-constraints via collaborative language agents. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10054–10082, Abu Dhabi, UAE. Association for Computational Linguistics.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Overview of Appendix Sections

- Appendix A: Prompts
- Appendix B: Hyper-parameter Details for Reproducibility
- Appendix C: Human Annotation Guidelines
- Appendix D: Detailed Analysis of Results
- Appendix E: Disruption Severity based Analysis of Results
- Appendix F: Rubric for LLM judging of disrupted travel plans
- Appendix G: Case Studies

A Prompts

This section details the carefully crafted prompts used in TripTide.

A.1 GPT-4o and Gemini 2.5 pro Prompt for Generating Disruption Queries

Instruction: You are an efficient AI agent to generate user query in natural language. You are given the following travel disruption information including disruption category, reason, disruption timestamp and disruption severity. Convert it into a travel disruption query by a user, which is a natural language query for the disruption information. The query should not include information like disruption severity that is unknown to a traveler and should be as natural as possible. The query must also ask for alternative options.

*****Example:*****

```
Disruption Information:
Day 1 - Lunch: Thames Street Oyster House
- Disruption Category: Restaurants
- Reason: The restaurant may be closed for a private media
launch event or local food festival buyout, making regular
dining unavailable for the public.
- Disruption Detection Timestamp: 1 day before (via restaurant
website or local event calendars).
- Disruption Severity: Step-level
-Confidence: Medium
```

Output:

```
Hey, I had planned to have lunch at Thames Street Oyster House on
Day 1, but I just found out today -- only a day before my visit --
that
it's closed due to a private event. Could you suggest a good
alternative?
```

*****Example Ends*****

```
Given disruption information:
{Disruption Information}
Output:
```

A.2 Prompt for GPT-4o based Revised Plan Generation

You are given a travel itinerary in JSON format annotation_plan, along with user details and disruption information disruption_info that affects the travel plan. The disruption has a severity level (step, day, or plan) indicating how much of the itinerary is impacted. The mitigation depends on the “Disruption Tolerance” level - Flexiventurer or Planbound. If the traveler is identified as “Planbound”, the scope of revision must strictly correspond to the disruption_severity. Specifically, for step-level disruptions, only the affected event should be modified; for day-level disruptions, modifications must be limited to the POIs scheduled for that particular day; and for plan-level disruptions, broader itinerary changes are permitted. In contrast, for “Flexiventurer” travelers, there is no constraint linking the revision scope to the disruption severity. Your task is to update the travel itinerary to accommodate the disruption with necessary changes, using the reference information reference_information provided to guide your modifications. Return the complete revised travel plan in the exact same JSON format as the original. You must acknowledge the disruption first and then proceed with appropriate revisions based on disruption severity and traveler’s disruption tolerance.

****** **Example** ******
Travel Itinerary and User Details:
{“idx”: 4, “JSON”: {“org”: “Tampa”, “dest”: “Bloomington”, “days”: 3, “visiting_city_number”: 1, “date”: ["2024-11-05”, “2024-11-06”, “2024-11-07”], “people_number”: 1, “local_constraint”: {“house rule”: null, “cuisine”: null, “room type”: null, “transportation”: null, “event”: null, “attraction”: null}, “budget”: 1650.0, “query”: null, “level”: “easy”}, “persona”: “Traveler Type: Adventure Seeker; Purpose of Travel: Nature; Spending Preference: Luxury Traveler; Location Preference: Forests/Wildlife”,"disruption tolerance”: “Planbound”, “plan”: [{“days”: 1, “current_city”: “From Tampa to Bloomington”, “transportation”: “Self-driving, from Tampa to Bloomington, Departure Time: 16:00, Arrival Time: 12:48”, “breakfast”: “-”, “attraction”: “Monroe Lake”, “lunch”: “Upland Brewing - Bloomington”, “dinner”: “Baxter’s American Grill”, “accommodation”: “Arcade House, Bloomington”, “event”: “-”, “point_of_interest_list”: “Arcade House, stay from 13:15 to 13:45, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away;Upland Brewing - Bloomington, lunch from 14:15 to 15:15, nearest transit: Rogers St & 11th St, 107.43m away;Monroe Lake, visit from 17:00 to 19:00, nearest transit: Rhorer Rd at the Clear Creek Crossing Shopping Center, 12621.65m away;Baxter’s American Grill, dinner from 20:00 to 21:30, nearest transit: Prow Rd @ Meadows Hospital, 247021.85m away;Arcade House, stay from 22:00 to 08:00, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away.”}, {“days”: 2, “current_city”: “Bloomington”, “transportation”: “-”, “breakfast”: “Osteria Rago”, “attraction”: “Miller Park Zoo; Constitution Trail; Tibetan Mongolian Buddhist Cultural Center”, “lunch”: “Runcible Spoon Cafe and Restaurant”, “dinner”: “Epiphany Farms Restaurant”, “accommodation”: “Arcade House, Bloomington”, “event”: “-”, “point_of_interest_list”: “Arcade House, stay from 08:00 to 08:30, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away;Osteria Rago, visit from 09:00 to 10:00, nearest transit: Dunn St and Kirkwood Ave, 78.59m away;Miller Park Zoo, visit from 10:30 to 12:30, nearest transit: Belle & Park Square, 252594.59m away;Runcible Spoon Cafe and Restaurant, visit from 13:00 to 14:30, nearest transit: Dunn St and Kirkwood Ave, 101.05m away;Constitution Trail, visit from 15:30 to 16:30, nearest transit: Prow Rd @ Meadows Hospital, 250910.58m away;Tibetan Mongolian Buddhist Cultural Center, visit from 17:30 to 19:30, nearest transit: Sare Rd @ Porto Flats Inbound, 1613.24m away;Epiphany Farms Restaurant, visit from 20:00 to 21:30, nearest transit: Prow Rd @ Meadows Hospital, 252068.66m away;Arcade House, stay from 22:00 to 07:00, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away.”}, {“days”: 3, “current_city”: “From Bloomington to Tampa”, “transportation”: “Taxi, from Bloomington to Tampa, Departure Time: 14:30, Arrival Time: 11:18”, “breakfast”: “The Chocolate Moose”, “attraction”: “Oliver Winery”, “lunch”: “Farm Bloomington”, “dinner”: “-”, “accommodation”: “-”, “event”: “-”, “point_of_interest_list”: “Arcade House, stay from 07:00 to 07:30, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away;The Chocolate Moose, visit from 08:00 to 09:00, nearest transit: Walnut St & Smith Ave @ Project School, 24.48m away;Oliver Winery, visit from 09:20 to 11:30, nearest transit: Acuff & Rosewood, 8560.79m away;Farm Bloomington, visit from 12:00 to 13:30, nearest transit: Washington St & Kirkwood Ave, 65.76m away.”}]}
Disruption Information:
Day 1- Accommodation: Arcade House, Bloomington
- Disruption Category: Accommodations
- Reason: Issues such as poor hygiene, missing amenities, or remote location may not meet luxury standards.
- Disruption Detection Timestamp: Upon arrival or 1 day before (confirmation with host/property).
- Disruption Severity: Step-level

Reference Information: { ... }

Complete Revised Travel Plan:
{“acknowledgement”: “Accommodation Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.”,“idx”: 4, “JSON”: {“org”: “Tampa”, “dest”: “Bloomington”, “days”: 3, “visiting_city_number”: 1, “date”: [“2024-11-05”, “2024-11-06”, “2024-11-07”], “people_number”: 1, “local_constraint”: {“house rule”: null, “cuisine”: null, “room type”: null, “transportation”: null, “event”: null, “attraction”: null}, “budget”: 1650.0, “query”: null, “level”: “easy"}, “persona”: “Traveler Type: Adventure Seeker; Purpose of Travel: Nature; Spending Preference: Luxury Traveler; Location Preference: Forests/Wildlife”, “disruption tolerance”: “Planbound”, “plan”: [{“days”: 1, “current_city”: “From Tampa to Bloomington”, “transportation”: “Self-driving, from Tampa to Bloomington, Departure Time: 16:00, Arrival Time: 12:48”, “breakfast”: “-”, “attraction”: “Monroe Lake”, “lunch”: “Upland Brewing - Bloomington”, “dinner”: “Baxter’s American Grill”, “accommodation”: “The Chateau Hotel and Conference Center, Bloomington”, “event”: “-”, “point_of_interest_list”: “The Chateau Hotel and Conference Center, check-in from 13:15 to 13:45,

address: 1621 Jumer Drive, Bloomington IL 61704, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away; Upland Brewing - Bloomington, lunch from 14:15 to 15:15, nearest transit: Rogers St & 11th St, 107.43m away; Monroe Lake, visit from 17:00 to 19:00, nearest transit: Rhorer Rd at the Clear Creek Crossing Shopping Center, 12621.65m away; Baxter's American Grill, dinner from 20:00 to 21:30, nearest transit: Prow Rd @ Meadows Hospital, 247021.85m away; The Chateau Hotel and Conference Center, overnight stay from 22:00 to 08:00, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away."},{"days": 2, "current_city": "Bloomington","transportation": "-","breakfast": "Osteria Rago", "attraction": "Miller Park Zoo; Constitution Trail; Tibetan Mongolian Buddhist Cultural Center", "lunch": "Runcible Spoon Cafe and Restaurant", "dinner": "Epiphany Farms Restaurant", "accommodation": "The Chateau Hotel and Conference Center, Bloomington", "event": "-", "point_of_interest_list": "The Chateau Hotel and Conference Center, stay from 08:00 to 08:30, address: 1621 Jumer Drive, Bloomington IL 61704, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away; Osteria Rago, visit from 09:00 to 10:00, nearest transit: Dunn St and Kirkwood Ave, 78.59m away; Miller Park Zoo, visit from 10:30 to 12:30, nearest transit: Belle & Park Square, 252594.59m away; Runcible Spoon Cafe and Restaurant, visit from 13:00 to 14:30, nearest transit: Dunn St and Kirkwood Ave, 101.05m away; Constitution Trail, visit from 15:30 to 16:30, nearest transit: Prow Rd @ Meadows Hospital, 250910.58m away; Tibetan Mongolian Buddhist Cultural Center, visit from 17:30 to 19:30, nearest transit: Sare Rd @ Porto Flats Inbound, 1613.24m away; Epiphany Farms Restaurant, visit from 20:00 to 21:30, nearest transit: Prow Rd @ Meadows Hospital, 252068.66m away; The Chateau Hotel and Conference Center, overnight stay from 22:00 to 07:00, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away."}, {"days": 3, "current_city": "From Bloomington to Tampa", "transportation": "Taxi, from Bloomington to Tampa, Departure Time: 14:30, Arrival Time: 11:18", "breakfast": "The Chocolate Moose", "attraction": "Oliver Winery", "lunch": "Farm Bloomington", "dinner": "-", "accommodation": "-", "event": "-", "point_of_interest_list": "The Chateau Hotel and Conference Center, check-out from 07:00 to 07:30, address: 1621 Jumer Drive, Bloomington IL 61704, nearest transit: Prow Rd @ Meadows Hospital, 251003.54m away; The Chocolate Moose, visit from 08:00 to 09:00, nearest transit: Walnut St & Smith Ave @ Project School, 24.48m away; Oliver Winery, visit from 09:20 to 11:30, nearest transit: Acuff & Rosewood, 8560.79m away; Farm Bloomington, visit from 12:00 to 13:30, nearest transit: Washington St & Kirkwood Ave, 65.76m away."}]}
****** **Example Ends** ******

Travel Itinerary and User Details:{annotation_plan}
Disruption Information:{disruption_info}
Reference Information :{reference_information}

Output the complete travel plan with acknowledgement and the modifications in the exact same JSON template as the original.
Your response must start with "Complete Revised Travel Plan"
Output: [{"Complete Revised Travel Plan": complete_revised_travel_plan}]

A.3 GPT-4o Prompt for Generating Possible Disruptions

Instruction: You are given a detailed travel itinerary in JSON format, which includes transportation, attractions, accommodations, meals, events, and the traveler's preferences. Based on this information, analyze the plan and identify potential types of disruptions or changes that are most likely to occur, classifying them into one or more of the following categories:
- Flights: Flight delays, cancellations, missed connections, baggage issues, strikes, or airport closures.
- Restaurants: Closure, changed hours, poor hygiene or food quality, no support for dietary needs, overcrowding, long wait times, price hikes due to events.
- Attractions: Maintenance/renovation, weather-related closures, government restrictions, ticket unavailability, ID requirements, overcrowding, specific day openings.
- Accommodations: Poor hygiene, remote/unsafe location, missing promised amenities.
- Miscellaneous: Event Cancellations, rescheduling, ticket issues, venue changes, weather or safety concerns.Forgotten documents, illness, fatigue, jet lag, bad weather.

Remember to give atleast three possible disruptions for each day of the plan. Example Format:

Most likely disruption categories:
- Flights: Justification...
- Restaurants: Justification...
...

Input:
{Insert the travel itinerary JSON here}

A.4 Prompt for Qwen2.5-7B-Instruct and Phi4-mini-Instruct based Revised Plan Generation

You are given a travel itinerary in JSON format, along with user details and a disruption information that affects the travel plan. The disruption has a severity level (step, day, or plan) indicating how much of the itinerary is impacted. The mitigation depends on the “Disruption Tolerance” level- Flexiventurer or Planbound. If the traveler is identified as “Planbound”, the scope of revision must strictly correspond to the disruption_severity. Specifically, for step-level disruptions, only the affected event should be modified; for day-level disruptions, modifications must be limited to the POIs scheduled for that particular day; and for plan-level disruptions, broader itinerary changes are permitted. In contrast, for “Flexiventurer” travelers, there is no constraint linking the revision scope to the disruption severity. Your task is to update the travel itinerary to accommodate the disruption with necessary changes, using the reference information provided to guide your modifications. Return the complete revised travel plan in the exact same JSON format as the original. You must acknowledge the disruption first and then proceed with appropriate revisions based on disruption severity and traveler’s disruption tolerance.

****** **Example** ******

Travel Itinerary and User Details:{“idx”: 25,“row_number”: 530,“JSON”: {“org”:“Green Bay”,“dest”:“Atlanta”,“days”: 3,“visiting_city_number”: 1,“date”: [“2024-11-20”,“2024-11-21”,“2024-11-22”],“people_number”: 2,“local_constraint”: {“house rule”: null,“cuisine”: [“Mexican”,“Italian”,“French”,“Mediterranean”],“room type”:“not shared room”,“transportation”: null,“event”: null,“attraction”: [“Shopping”,“Concerts & Shows”]},“budget”: 1650.0,“query”: null,“level”:“hard”},“persona”:“Traveler Type: Adventure Seeker; Purpose of Travel: Cultural Exploration; Spending Preference: Luxury Traveler; Location Preference: Cities”,“disruption tolerance”:“Planbound”,“plan”: [{“days”: 1,“current_city”:“from Green Bay to Atlanta”,“transportation”:“Flight Number: F3526412, from Green Bay to Atlanta, Departure Time: 05:32, Arrival Time: 08:50”,“breakfast”:“St. Cecilia, Atlanta”,“attraction”:“High Museum of Art, Atlanta;World of Coca-Cola,Atlanta; Fox Theatre, Atlanta”,“lunch”:“La Tavola Trattoria, Atlanta”,“dinner”:“Nikolai’s Roof, Atlanta”,“accommodation”:“Chic, Clean, City Loft Getaway, Atlanta”,“event”:“-”,“point_of_interest_list”:“Chic, Clean, City Loft Getaway, stay from 09:30 to 10:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;St. Cecilia, visit from 10:30 to 11:00, nearest transit: PEACHTREE RD NE @ OAK VALLEY RD NE, 63.53m away;High Museum of Art, visit from 12:00 to 14:00, nearest transit: PEACHTREE ST NE @ 16TH ST NE, 81.08m away;La Tavola Trattoria, visit from 15:00 to 15:30, nearest transit: VIRGINIA AVE NE @ TODD RD NE, 27.36m away;World of Coca-Cola, visit from 16:00 to 17:30, nearest transit: IVAN ALLEN JR BLVD NW @ LOVEJOY ST, 249.77m away;Fox Theatre, visit from 18:00 to 20:45, nearest transit: PEACHTREE ST NE @ PONCE DE LEON AVE, 41.68m away;Nikolai’s Roof, visit from 21:30 to 22:30, nearest transit: RALPH MCGILL BLVD NE @ COURTLAND ST, 309.92m away;Chic, Clean, City Loft Getaway, stay from 23:00 to 07:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away.”}, {“days”: 2,“current_city”:“Atlanta”,“transportation”:“-”,“breakfast”:“Home Grown, Atlanta”,“attraction”:“Zoo Atlanta, Atlanta; National Center for Civil and Human Rights, Atlanta”,“lunch”:“No Mas! Cantina, Atlanta”,“dinner”:“Cafe Agora, Atlanta”,“accommodation”:“Chic, Clean, City Loft Getaway, Atlanta”,“event”:“-”,“point_of_interest_list”:“Chic, Clean, City Loft Getaway, stay from 07:00 to 07:30, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;Home Grown, visit from 08:00 to 09:00, nearest transit: MEMORIAL DR SE @ GIBSON ST SE, 49.17m away;Zoo Atlanta, visit from 09:30 to 13:00, nearest transit: CHEROKEE AVE @ GRANT PARK PL, 195.14m away;No Mas! Cantina, visit from 13:30 to 15:00, nearest transit: PETERS ST SW @ HAYNES ST SW, 123.79m away;National Center for Civil and Human Rights, visit from 15:30 to 18:30, nearest transit: IVAN ALLEN JR BLVD NW @ LOVEJOY ST, 78.15m away;Cafe Agora, visit from 20:00 to 21:30, nearest transit:PEACHTREE ST @ PEACHTREE PL , 29.31m away;Chic, Clean, City Loft Getaway, stay from 22:00 to 07:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away.”}, {“days”: 3,“current_city”:“from Atlanta to Green Bay”,“transportation”:“Flight Number: F0889249, from Atlanta to Green Bay, Departure Time: 17:56, Arrival Time: 18:55”,“breakfast”:“South City Kitchen Midtown, Atlanta”,“attraction”:“Piedmont Park, Atlanta; Lenox Square, Atlanta”,“lunch”:“Aviva by Kameel, Atlanta”,“dinner”:“-”,“accommodation”:“-”,“event”:“-”,“point_of_interest_list”:“Chic, Clean, City Loft Getaway, stay from 07:00 to 08:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;South City Kitchen Midtown, visit from 09:00 to 10:00, nearest transit: PEACHTREE ST @ 14TH ST NE, 128.52m away;Piedmont Park, visit from 10:30 to 13:30, nearest transit: PIEDMONT AVE NE @ PRADO, 407.10m away;Aviva by Kameel, visit from 14:05 to 15:00, nearest transit: PEACHTREE ST @ ANDREW YOUNG INTL BLVD (PEACHTREE CTR STN), 52.29m away;Lenox Square, visit from 15:30 to 17:00, nearest transit: PEACHTREE RD NE @ LENOX PKWY, 230.44m away.”}]}

Disruption Query: I was thinking of doing some shopping at Lenox Square in Atlanta on Day 3, but I just heard there might be temporary closures due to a security drill or incident. Do you know if it’s still open? If not, could you point me to another good shopping spot nearby?

Reference Information: {...}

Complete Revised Travel Plan:
acknowledgement: Attractions Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.
Day 1:

Current City: from Green Bay to Atlanta
Transportation: Flight Number: F3526412, from Green Bay to Atlanta, Departure Time: 05:32, Arrival Time: 08:50
Breakfast: St. Cecilia, Atlanta
Attraction: High Museum of Art, Atlanta;World of Coca-Cola,Atlanta; Fox Theatre, Atlanta
Lunch: La Tavola Trattoria, Atlanta
Dinner: Nikolai's Roof, Atlanta
Accommodation: Chic, Clean, City Loft Getaway, Atlanta
Event: -
Point of Interest List: Chic, Clean, City Loft Getaway, stay from 09:30 to 10:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;St. Cecilia, visit from 10:30 to 11:00, nearest transit: PEACHTREE RD NE @ OAK VALLEY RD NE, 63.53m away;High Museum of Art, visit from 12:00 to 14:00, nearest transit: PEACHTREE ST NE @ 16TH ST NE, 81.08m away;La Tavola Trattoria, visit from 15:00 to 15:30, nearest transit: VIRGINIA AVE NE @ TODD RD NE, 27.36m away;World of Coca-Cola, visit from 16:00 to 17:30, nearest transit: IVAN ALLEN JR BLVD NW @ LOVEJOY ST, 249.77m away;Fox Theatre, visit from 18:00 to 20:45, nearest transit: PEACHTREE ST NE @ PONCE DE LEON AVE, 41.68m away;Nikolai's Roof, visit from 21:30 to 22:30, nearest transit: RALPH MCGILL BLVD NE @ COURTLAND ST, 309.92m away;Chic, Clean, City Loft Getaway, stay from 23:00 to 07:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away.

Day 2:
Current City: Atlanta
Transportation: -
Breakfast: Home Grown, Atlanta
Attraction: Zoo Atlanta, Atlanta; National Center for Civil and Human Rights, Atlanta
Lunch: No Mas! Cantina, Atlanta
Dinner: Cafe Agora, Atlanta
Accommodation: Chic, Clean, City Loft Getaway, Atlanta
Event: -
Point of Interest List: Chic, Clean, City Loft Getaway, stay from 07:00 to 07:30, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;Home Grown, visit from 08:00 to 09:00, nearest transit: MEMORIAL DR SE @ GIBSON ST SE, 49.17m away;Zoo Atlanta, visit from 09:30 to 13:00, nearest transit: CHEROKEE AVE @ GRANT PARK PL, 195.14m away;No Mas! Cantina, visit from 13:30 to 15:00, nearest transit: PETERS ST SW @ HAYNES ST SW, 123.79m away;National Center for Civil and Human Rights, visit from 15:30 to 18:30, nearest transit: IVAN ALLEN JR BLVD NW @ LOVEJOY ST, 78.15m away;Cafe Agora, visit from 20:00 to 21:30, nearest transit:PEACHTREE ST @ PEACHTREE PL , 29.31m away;Chic, Clean, City Loft Getaway, stay from 22:00 to 07:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away.

Day 3:
Current City: from Atlanta to Green Bay
Transportation: Flight Number: F0889249, from Atlanta to Green Bay, Departure Time: 17:56, Arrival Time: 18:55
Breakfast: South City Kitchen Midtown, Atlanta
Attraction: Piedmont Park, Atlanta; Ponce City Market, Atlanta
Lunch: Aviva by Kameel, Atlanta
Dinner: -
Accommodation: -
Event: -
Point of Interest List: Chic, Clean, City Loft Getaway, stay from 07:00 to 08:00, nearest transit: PONCE DE LEON AVE @ LULLWATER RD NE, 192.63m away;South City Kitchen Midtown, visit from 09:00 to 10:00, nearest transit: PEACHTREE ST @ 14TH ST NE, 128.52m away;Piedmont Park, visit from 10:30 to 13:30, nearest transit: PIEDMONT AVE NE @ PRADO, 407.10m away;Aviva by Kameel, visit from 14:05 to 15:00, nearest transit: PEACHTREE ST @ ANDREW YOUNG INTL BLVD (PEACHTREE CTR STN), 52.29m away;Ponce City Market, visit from 15:30 to 17:00, nearest transit: PONCE DE LEON AVE NE @ GLEN IRIS DR NE, 150.00m away.

****** **Example Ends** ******
Travel Itinerary and User Details:{text}
Disruption Query:{query}
Reference Information :{reference_information}
Complete Revised Travel Plan"

844

A.5 Prompt for Llama-3.1-8B-Instruct used as LLM-as-Judge

JUDGE_PROMPT = """You are a penalty-focused, highly critical evaluator of travel plan revisions. You are given:
1. The original travel plan
2. The disruption description
3. The revised plan

Traveler Type Definitions
- **Planbound** = Must follow the original plan exactly, unless absolutely forced by disruption. *Step-level disruption → only that step may change. Day-level disruption → only that day may change. Plan-level disruption → broader edits allowed. Any change outside this scope = automatic penalty (max score 2–3).*
- **Flexiventurer** = Allows more flexibility, but even here, unrelated or unnecessary changes = penalty. — **### Scoring Rubric**
- ***5 (Excellent)***
Disruption handled clearly and effectively:
• Disrupted POI replaced/removed properly.
• Only necessary changes made (no extras).
• Traveler constraints fully respected.
• Plan is coherent and realistic.
→ Use 5 if handling is perfect OR very close to perfect.
- ***4 (Good)***
Disruption handled correctly, but with minor flaws:
• Traveler = Flexiventurer → a small unnecessary edit (e.g., time shift, tiny change).
• Traveler = Planbound → disruption fixed properly, but very small deviation outside scope.
→ 4 means "handled well but not flawless."
- ***3 (Average)***
Disruption addressed, but with noticeable problems:
• Planbound → multiple edits outside allowed scope.
• Flexiventurer → unnecessary changes beyond what was required.
• Some sequencing / feasibility issues (but plan still makes sense overall).
→ 3 = "okay, but sloppy or constraint-violating."
- ***2 (Poor)***
Disruption weakly or superficially handled:
• Only acknowledgement text without real itinerary fix.
• Disrupted POI still present but disruption noted.
• Major incoherence in plan (e.g., order, timing).
→ 2 = "attempt made, but it fails in practice."
- ***1 (Very Poor)***
Disruption ignored entirely:
• Revised plan identical to original, OR
• Disrupted POI remains unchanged with no mitigation.
• Changes are irrelevant to disruption.
→ 1 = "clear evidence disruption not handled at all."

**Default to lower scores if there is *any doubt*.
** Scores 4 and 5 should be *rare exceptions*, only for near-perfect plans.
—
Examples
Case A (Very Poor – 1)
Original Plan: Dinner at "Bone's Restaurant, Atlanta". Disruption: Restaurant closed due to private event. Revised Plan: Still lists "Bone's Restaurant" for dinner. → Score = 1 ("Disruption ignored, POI unchanged.")
Case B (Poor – 2)
Original Plan: Day 3 visit to "Lenox Square, Atlanta". Disruption: Mall closed due to security incident. Revised Plan: Starts with acknowledgement but still lists Lenox Square in itinerary. → Score = 2 ("Disruption noted but not properly mitigated; POI remains.")
—
Now evaluate: Original Plan: {initial_plan}
Disruption: {disruption_info}
Revised Plan: {mitigated_plan}

Answer ONLY in JSON:
{{ "score": X,
"explanation": "<explicit reasoning, listing penalties applied and why>" }} """

B Hyper-parameter Details for Reproducibility

In this paper, we have used GPT-4o, Qwen2.5-7B-Instruct, and Phi4-mini-Instruct LLM models to generate the revised travel plan based on our dataset. The hyperparameter details of the three models are listed below:

Parameter	GPT-4o	Qwen2.5-7B instruct	Phi-4-mini instruct
Model Source	OpenAI	Alibaba (via HuggingFace)	Microsoft (via HuggingFace)
Model Size	Unknown	7B	2.7B
Temperature	0.0	0.0	0.0
Top-p	1.0	1.0	1.0
Max Tokens	10,000	3072	4096
Frequency Penalty	0.0	0.0	0.0
Presence Penalty	0.0	0.0	0.0
System Prompt	Github Repo	Github Repo	Github Repo
Version	May 2024	Sep 2024	Feb 2025

Table 4: Model Inference Hyperparameters

Server Configuration Details. All local inference experiments were conducted on a high-performance server with the following system specifications:

- **Operating System:** Ubuntu 24.04.2 LTS
- **CPU:** Dual AMD EPYC 9474F (2 × 48-Core, 192 threads total)
- **GPU:** NVIDIA L40, 46 GB VRAM
- **CUDA Version:** 12.9
- **NVIDIA Driver Version:** 575.64.03
- **Python Version:** 3.10
- **Inference Libraries:**
  - `transformers==4.53.2`
  - `torch==2.5.1`
  - `accelerate==1.9.0`
- **Memory:** 46 GB GPU memory, 1TiB system RAM
- **Model Hosting:**
  - GPT-4o: Accessed via OpenAI API
  - Qwen2.5-7B-Instruct: Inferred locally using Hugging Face model `Qwen/Qwen2.5-7B-Instruct`
  - Phi-4-mini-Instruct: Inferred locally using Hugging Face model `microsoft/Phi-4-mini-instruct`

#	Annotation Guidelines
1	The objective is to generate an alternative feasible plan following the occurrence of a disruption.
2.	Input: Current plan, disruption, disruption severity, disruption_tolerance Output: alternative feasible plan.
3	The alternative plan should introduce only the minimal necessary modifications, and the updated list of Points of Interest (POIs) must not include any fabricated or unverified information. The feasibility will be checked by us using scripts.
4	The annotators have to keep in mind that the disruption mitigation process depends on two factors: both Disruption_severity and Disruption_tolerance.Refer Table 1.
5	The extent of changes should correspond to the disruption_severity: **Step-level:** disruptions should involve the least amount of change, changing only the disrupted event. **Day-level:** disruptions may require moderate adjustment, meaning modifying the disrupted day's itinerary only. **Plan-level:** disruptions may necessitate more extensive revisions. The annotators are free to change the entire plan, if necessary.
6	Annotators must consider the traveler's disruption_tolerance attribute (Flexiventurer / Planbound) when modifying the plan, and should provide clear remarks explaining the adjustments made.
7	The travelers are categorized based on their "disruption_tolerance": **Flexiventurer:** Open to last-minute modifications, such as rearranging the itinerary by substituting POIs from subsequent days to optimize the overall experience. **Planbound:** Prefer to adhere strictly to the original itinerary and are reluctant to accept any significant changes.
8	The annotation process must consider the traveler's classification. If the traveler is identified as "Planbound", the scope of revision must strictly correspond to the disruption_severity. Specifically, for step-level disruptions, only the affected event should be modified; for day-level disruptions, modifications must be limited to the POIs scheduled for that particular day; and for plan-level disruptions, broader itinerary changes are permitted. In contrast, for "Flexiventurer" travelers, there is no constraint linking the revision scope to the disruption severity. Annotators are granted the flexibility to revise the itinerary as deemed appropriate. However, any such modifications must be accompanied by a clear justification in the "Remarks" column."
9	While generating the revised POI list, annotators must preserve the original user persona characteristics. For example: A laidback traveler may prefer a schedule with 1-2 attractions per day, even when additional options are available. An economical traveler would prioritize budget-friendly options over more expensive alternatives.
10	Annotators are required to provide truthful and contextually appropriate responses by generating a revised travel plan that differs from the original itinerary, effectively addressing and mitigating the impact of the identified disruption
11.	Annotators should apply their best judgment to ensure that the updated plans are both practical and realistic. Any major decision-making rationale must be explicitly documented in the Remarks section.

Table 5: Guidelines for Annotation of Revised Travel Plans and Remarks

C Human Annotation Guidelines

Guidelines for Annotation Annotators were instructed to manually identify potential day-wise disruptions for each travel plan. The goal was to sample disruptions in a balanced manner across key categories: *Accommodation*, *Attractions*, *Restaurants*, *Transport*, and *Miscellaneous*. Once disruptions were sampled, annotators were required to generate a corresponding Revised Plan. This plan had to be created using the *Disruption Query*, the *Initial Plan*, and the *Reference Information*. The annotated plan was expected to strictly adhere to the specified *User Persona*, *Disruption Severity*, and the *Traveler's Tolerance Level*. Please refer to Table 5 for detailed annotation guidelines.

Annotator Demographics The demographic distribution of annotators reflects a diverse range of educational backgrounds and experience levels. English proficiency distribution of our graduate student annotators is as follows: 8 years (6.7%), 12 years (26.7%), 16 years (53.3%), 20 years (13.3%). This shows that the majority of annotators have received 12 to 20 years of formal English education, indicating a high level of language proficiency. Age distribution of our graduate student annotators is as follows: 18 years (3.3%), 19 years (20.0%), 20 years (26.7%), 21 years (33.3%), 22 years (6.7%), 23 years (10.0%). Thus, the age distribution is centered around 20 to 25 years, suggesting that most annotators are young graduate students in early-to-mid adulthood. Gender representation (Male: 66.7%, Female: 33.3%) also shows a balanced participation across genders among the 30 graduate student annotators. These demographics suggest that our annotators possess strong language and reasoning abilities, along with the cognitive maturity required to effectively evaluate and annotate revised travel plans.

D Detailed Analysis of Results

We report the performance[7] of GPT-4o and Qwen2.5-7B-Instruct and Phi-4-mini Instruct across initial and revised itineraries using *Preservation of Intent* metrics, *Adaptability* metrics (*semantic*, *spatial*, *sequential*), and *Responsiveness* in Table 3 across three plan durations: 3-day, 5-day, and 7-day trips.

Preservation of Intent. Both GPT-4o and Qwen2.5-7B-Instruct achieve high delivery rates ($\geq$99.69%) for 3-day and 5-day plans, with a slight drop observed for Qwen2.5-7B-Instruct at 7-day (97.54%), whereas Phi-4 mini struggles to deliver for shorter travel plans but provides a perfect score for 7-day (100.00%). Across commonsense (CPR) and hard constraint (HCPR) metrics, GPT-4o consistently achieves moderate CPR and HCPR micro scores, with the best performance on 5-day plans. Qwen-7B-Instruct shows comparable or slightly higher commonsense compliance for shorter durations, but its hard constraint adherence deteriorates sharply in 7-day plans (HCPR micro drops to 25.16%). On the other hand, for Phi-4 mini Instruct, CPR is uneven, i.e., micro improves with duration (39.75$\rightarrow$68.81$\rightarrow$73.43) while macro collapses (18.68$\rightarrow$6.18$\rightarrow$3.06); HCPR stays low and declines (micro 20.36$\rightarrow$18.89$\rightarrow$7.67, macro 24.49$\rightarrow$13.28$\rightarrow$7.96). This suggests that Qwen2.5-7B-Instruct and Phi4-mini-Instruct's constraint reasoning ability diminishes with increasing plan complexity, while GPT-4o maintains more stable performance. In terms of final pass rates, GPT-4o leads overall: peaking on 5-day plans at 41.98% (29.37% for 3-day, 32.54% for 7-day). Qwen2.5-7B-Instruct remains around $\sim$30% on 3- and 5-day plans (29.98% and 29.12%) but drops markedly to 16.27% on 7-day, while Phi-4 mini Instruct is consistently low (18.26%$\rightarrow$5.87%$\rightarrow$3.06%). These trends emphasize GPT-4o's, Qwen2.5-7B-Instruct's and Phi-4-mini Instruct's reliability in producing plans, while also underscoring the need for improved mechanisms to ensure comprehensive constraint satisfaction.

Semantic Adaptability ($A_{\mathbf{sem}}$). GPT-4o maintains strong semantic fidelity across all plan durations. The differences between initial and revised semantic scores are low (lower is better): 0.34 (3-day), 0.02 (5-day), and 0.20 (7-day), indicating minimal semantic drift post-revision. The near-zero shift for 5-day plans suggests a precise understanding of user intent, even under revision constraints. Qwen2.5-7B-Instruct also achieves consistently low semantic drift across all durations: 0.67 (3-day), 0.04 (5-day), and 1.24 (7-day). These results indicate strong alignment with user personas during plan revisions, on par with GPT-4o. The slightly higher shift for 7-day plans could be attributed to inherent possibility of higher drifts in re-planning when facing disruptions in longer itineraries. While Phi4-mini-Instruct results show decreasing semantic coherence between original and revised plan, where the semantic scores are as high as 30.34%. This indicates that the Phi-4-mini struggles to understand the user's intent under disruptions.

Spatial Adaptability ($A_{\mathbf{spa}}$). For GPT-4o, the spatial score remains low (lower is better) across revisions: 1.22 (3-day), 0.66 (5-day), and 0.30 (7-day), indicating minimal spatial drift between the original and revised plans, with the smallest drift for longer (7-day) itineraries. These indicate that GPT-4o consistently restructures plans to group spatially adjacent entities after disruption. The smaller drift for longer duration plans supports the hypothesis that the model leverages spatial flexibility more readily in high-duration settings. On the other hand, Qwen2.5-7B-Instruct leads

[7]We evaluate the models on the entire dataset due to the absence of a predefined train-test split. These evaluations are diagnostic in nature and are intended to inform design insights rather than generalization performance.

to higher spatial drift across all durations: 5.94 (3-day), 1.25 (5-day), and 2.10 (7-day). These results indicate that Qwen2.5-7B-Instruct leads to significant spatial score changes compared to GPT-4o. Phi-4-mini-Instruct shows modest drift for shorter plans (3-day: 3.32; 5-day: 1.45) but fails to maintain spatial coherence for 7-day, with a score of 27.78, indicating difficulty scaling to longer itineraries.

Sequential Adaptability ($A_{\mathbf{seq}}$). Sequential adaptability measures the average change in PoI ordering across days, based on normalized edit distance between the initial and revised plans. A lower A_{seq} implies lower disruption to the temporal structure. As presented in Table 3, GPT-4o exhibits consistently low sequential scores (lower is better): 10.89 (3-day), 1.86 (5-day), and 5.58 (7-day). These scores indicate minimal reordering of PoIs post-disruption, particularly in 5-day itineraries, suggesting that GPT-4o prioritizes original sequence structure while performing disruption constraint satisfaction. In contrast, Qwen2.5-7B-Instruct shows significantly higher A_{seq} scores: 56.26 (3-day), 28.46 (5-day), and 24.04 (7-day), indicating much weaker alignment with the original temporal order. While the score decreases with plan length, due to accumulating complexity, Qwen2.5-7B-Instruct remains relatively less sequence-preserving than GPT-4o. Similarly, Phi-4-mini instruct also shows higher sequential scores: 26.64 (3-day), 30.13 (5-day) and 42.74 (7-day). The longer the plan horizon, the higher the sequential score. This indicates the model performs poorly in maintaining the sequentiality of the POI list when the plan horizon increases. This contrast highlights a key trade-off: Qwen2.5-7B-Instruct's and Phi-4-mini Instruct's aggressive re-optimization often disrupts sequence, while GPT-4o adopts a more conservative, order-aware strategy.

Responsiveness Rate. Responsiveness, defined as the proportion of disruption-handled plans, drops from 89.53% (3-day) to 81.79% (5-day) and further to 79.82% (7-day). This monotonic decline suggests that GPT-4o faces increasing difficulty as plan complexity and temporal scope increase. Despite strong semantic and spatial preservation, disruption resolution becomes harder to maintain across extended itineraries. Qwen2.5-7B-Instruct follows the same downward trend with increasing duration from 3-day to 5-day but an increase for 7-day: 97.67% (3-day), 82.02% (5-day), and 88.67% (7-day). Interestingly, Qwen2.5-7B-Instruct achieves a higher responsiveness rate in longer plans compared to GPT-4o, despite more aggressive spatial edits, which indicates better disruption handling at scale. The phi-4-mini instruct model gives again surprising results while checking for mitigation nature of the revised plan. The models gives 100% responsiveness rate for 7-day travel planning, while the rate for 3-day and 5-day are low.

E Disruption Severity based Analysis of Results

In this section we analyze how the models perform based on the disruption severity (Step, Day or Plan) irrespective of travel duration. Results are presented in Tables 6 and 7.

Preservation of Intent: Table 6 shows that GPT-4o maintains a perfect delivery rate across most disruption levels, with strong performance in both commonsense (CPR) and hard constraint (HCPR) pass rates, particularly excelling at Step and Plan-level disruptions. Qwen-2.5 7B instruct also sustains a high delivery rate, but its HCPR scores remain notably lower, especially under Step-level disruptions, indicating weaker handling of strict constraints. While the models show stable CPR performance, GPT-4o consistently achieves better constraint satisfaction overall indicating the model's strength to capture the intent across all the disruption severity levels.

Semantic Adaptability ($A_{\mathbf{sem}}$): Table 7 reflects that GPT-4o achieves the lowest semantic score at Step-level (0.05), indicating strong preservation of original plan intent under fine-grained disruptions. However, it worsens slightly at Day (0.41) and Plan-level (0.17). In contrast, Qwen2.5-7B-Instruct maintains slightly higher and relatively constant semantic scores (0.40-1.46), suggesting that its plan revisions diverge more from the original semantic context, particularly under Day-level disruptions.

Spatial Adaptability ($A_{\mathbf{spa}}$): GPT-4o exhibits stable spatial consistency across all levels, with its best score at Day-level (0.21). Qwen2.5-7B-Instruct instruct struggles at Step-level with a high score of 5.49, implying poor location coherence, but performs best at Day-level (0.01). This reflects the fact that Qwen2.5-7B-Instruct adapts better when the disruption granularity is coarser.

Model Name	Disruption Severity	Delivery Rate↑	CPR↑ Micro	CPR↑ Macro	HCPR↑ Micro	HCPR↑ Macro	Final Pass Rate↑
GPT-4o	Step-Level	100.00	**90.29**	32.46	47.68	46.56	31.15
	Day- level	100.00	89.94	**33.78**	54.15	**59.68**	**33.45**
	Plan-level	78.69	70.49	30.17	**56.09**	55.74	29.51
Qwen2.5-7B-Instruct	Step Level	99.65	86.98	24.49	42.47	44.22	24.49
	Day- level	99.33	86.41	25.43	44.23	45.43	24.07
	Plan-level	98.71	85.86	28.14	51.54	49.36	28.14

Table 6: Preservation of user intent scores for different Models based on Disruption Severity level.

Sequential Adaptability (A_{seq}): GPT-4o shows excellent temporal consistency at Day and Plan-levels (as low as 5), but much higher at Step-level (82.36), indicating difficulty in preserving activity sequences during fine-grained changes. On the other hand, Qwen2.5-7B-Instruct performs moderately with a score of nearly 30, suggesting GPT-4o performs better when maintaining the sequentiality of the POI list with respect to the original plan for day-level and plan-level disruptions.

Responsiveness rate: Qwen2.5-7B-Instruct outperforms GPT-4o while checking the mitigation rate of the models between the final revised and initial plan. We can observe from Table 7 that for Qwen2.5-7B-Instruct the rate is nearly 100% while GPT-4o performance degrades as the severity level increases.

Summary. In this paper we analyzed capacity of GPT-4o, Qwen2.5-7B-Instruct to handle the disruptions for both when the dataset is divided among Travel Durations (3, 5, 7-days) and Disruption Severity (Step, Day, and Plan Levels). From Table 6 and Table 7, we can infer that, in the day-wise setting, GPT-4o maintains high sequential consistency and improving spatial scores with longer plans, showing stable adaptability. Qwen2.5-7B-Instruct and Phi-4-mini-Instruct achieve higher responsiveness but struggle with semantic and spatial coherence, especially for 3-day plans. In the severity-wise setting, GPT-4o excels in semantic alignment under step-level disruptions but drops in responsiveness. Meanwhile, Qwen2.5-7B-Instruct retain high responsiveness across severities but shows poor sequential alignment at finer-grained disruptions.

F Rubric for LLM judging of disrupted travel plans

Following the *LLM-as-a-Judge* paradigm (Zheng et al., 2023), beyond the automated metrics in Section 4, we further assess revised plans generated by GPT-4o using an independent LLM. We choose

Model Name	Disruption Severity	Semantic Score (A_{sem}) ↓	Spatial Score (A_{spa}) ↓	Sequential Score (A_{seq}) ↓	Responsiveness Rate ↑
GPT-4o	Step-level	**0.05**	0.65	82.36	94.11
	Day-level	0.41	0.21	4.98	78.84
	Plan-level	0.17	1.41	**4.55**	76.76
Qwen2.5-7B-instruct	Step-level	0.40	5.49	34.06	96.39
	Day-level	1.46	**0.01**	32.54	**99.67**
	Plan-level	0.41	1.31	31.42	99.57

Table 7: Adaptability (A_{sem}, A_{spa}, A_{seq}) and Responsiveness scores (%) for different models based on Disruption Severity level.

GPT-4o as it achieves the best average performance across metrics (Table 3). For each query, the judge is provided with the original plan, disruption details, and the revised plan. The revised itinerary is rated on a 1–5 scale, where 1 denotes minimal or ineffective revision and 5 indicates complete resolution of the disruption. The scoring rubric is shown in Table 8.

G Case Studies

In this section we showcase a few of the examples that highlight the model's strengths and weaknesses while handling the disruptions.

Rating	Description
5 — Excellent	Disruption fully fixed; only necessary edits; traveler constraints and persona respected; plan coherent and realistic.
4 — Good	Handled correctly but minor flaws like tiny unnecessary change (Flexiventurer) or very small out-of-scope tweak (Planbound).
3 — Average	Fixed but sloppy: extra edits beyond scope (Planbound) or unnecessary changes (Flexiventurer); minor timing/sequence issues.
2 — Poor	Superficial fix: only an acknowledgement; disrupted POI still present; major ordering/timing incoherence.
1 — Very Poor	Disruption ignored: plan unchanged or disrupted POI untouched; edits irrelevant to the disruption.

Table 8: Rubric for LLM judging of disrupted travel plans

G.1 Example showing how GPT-4o effectively mitigated the disruption while maintaining the Semantic and Traveler tolerance level.

Initial Plan {"idx": 27, "row_number": "532-1", "JSON": {"org": "Miami", "dest": "Baltimore", "days": 3, "visiting_city_number": 1, "date": ["2024-11-03", "2024-11-04", "2024-11-05"], "people_number": 2, "local_constraint": {"house rule": null, "cuisine": null, "room type": null, "transportation": null, "event": ["Arts & Theatre"], "attraction": null}, "budget": 1800.0, "query": null, "level": "medium"}, "persona": "Traveler Type: Adventure Seeker; Purpose of Travel: Adventure; Spending Preference: Luxury Traveler; Location Preference: Mountains", "disruption tolerance": "Planbound", "plan": [{"days": 1, "current_city": "from Miami to Baltimore", ... },
{"days": 2, "current_city": "Baltimore", ... },
{"days": 3, "current_city": "from Baltimore to Miami", "transportation": "Flight Number: F2123935, from Baltimore to Miami, Departure Time: 12:11, Arrival Time: 14:31", "breakfast": "Blue Moon Cafe, Baltimore", "attraction": "Fort McHenry National Monument And Historic Shrine, Baltimore", "lunch": "-", "dinner": "-", "accommodation": "-", "event": "-", "point_of_interest_list": "Luxe Houseboat: Stunning Inner Harbor 360 Views, stay from 07:00 to 07:30, nearest transit: SHOT TOWER METRO West Entrance, 841.96m away;Blue Moon Cafe, breakfast from 08:00 to 09:00, nearest transit: SHOT TOWER STATION (METRO) sb, 1161.67m away;Fort McHenry National Monument And Historic Shrine, visit from 09:30 to 11:30, nearest transit: SHOT TOWER STATION (METRO) sb, 3547.78m away."}]}
Disruption Info: Day 3 Attraction: Fort McHenry National Monument
Disruption Category: Attractions
Reason: Historic sites often have limited morning hours, and visit ends very close to flight check-in.
Disruption Detection Timestamp: 1 day before
Disruption Severity: Day-level
Confidence Level: Medium

Revised Plan {"acknowledgement": "Attractions Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.", "idx": 27, "row_number": "532-1", "JSON": {"org": "Miami", "dest": "Baltimore", "days": 3, "visiting_city_number": 1, "date": ["2024-11-03", "2024-11-04", "2024-11-05"], "people_number": 2, "local_constraint": {"house rule": null, "cuisine": null, "room type": null, "transportation": null, "event": ["Arts & Theatre"], "attraction": null}, "budget": 1800.0, "query": null, "level": "medium"}, "persona": "Traveler Type: Adventure Seeker; Purpose of Travel: Adventure; Spending Preference: Luxury Traveler; Location Preference: Mountains", "disruption tolerance": "Planbound", "plan": [{"days": 1, "current_city": "from Miami to Baltimore", ...},

{"days": 2, "current_city": "Baltimore", ... },
{"days": 3, "current_city": "from Baltimore to Miami", "transportation": "Flight Number: F2123935, from Baltimore to Miami, Departure Time: 12:11, Arrival Time: 14:31", "breakfast": "Blue Moon Cafe, Baltimore", "attraction": "Inner Harbor, Baltimore", "lunch": "-", "dinner": "-", "accommodation": "-", "event": "-", "point_of_interest_list": "Luxe Houseboat: Stunning Inner Harbor 360 Views, stay from 07:00 to 07:30, nearest transit: SHOT TOWER METRO West Entrance, 841.96m away;Blue Moon Cafe, breakfast from 08:00 to 09:00, nearest transit: SHOT TOWER STATION (METRO) sb, 1161.67m away;Inner Harbor, visit from 09:30 to 10:30, nearest transit: SHOT TOWER STATION (METRO) sb, 3547.78m away."}]}

Analysis: The model mitigated the Day 3 disruption which is a *Day-level disruption* with another apprpriate attraction type. The model was also able to capture the tolerance level of traveler (*Planbound)* as all other day's POI list remains same.

G.2 Example citing the GPT-4o model's struggles to mitigate the Disruption for 5-day travel Planning

Initial Plan {"idx": 177, "row_number": "65-1", "JSON": {"org": "Lynchburg", "dest": "North Carolina", "days": 5, "visiting_city_number": 2, "date": ["2024-11-02", "2024-11-03", "2024-11-04", "2024-11-05", "2024-11-06"], "people_number": 1, "local_constraint": {"house rule": "parties", "cuisine": null, "room type": "entire home", "transportation": null, "event": null, "attraction": "Nature & Parks"}, "budget": 1800.0, "query": null, "level": "hard"}, "persona": "Traveler Type: Laidback Traveler; Purpose of Travel: Nature; Spending Preference: Economical Traveler; Location Preference: Beaches", "disruption tolerance": "Flexiventurer", "plan": [{"days": 1, "current_city": "from Lynchburg to Greensboro", ...}
{"days": 2, "current_city": "Greensboro", ... }, ...

{"days": 4, "current_city": "Charlotte", "transportation": "-", "breakfast": "The Fig Tree Restaurant, Charlotte", "attraction": "Mint Museum Uptown, Charlotte; The Bank Of America Stadium, Charlotte", "lunch": "Princess Charlotte, Charlotte", "dinner": "Cabo Fish Taco Baja Seagrill, Charlotte", "accommodation": "Tippah Treehouse Retreat, Charlotte", "event": "-", "point_of_interest_list": " Tippah Treehouse Retreat, stay from 09:00 to 09:30, nearest transit: The Plaza & Marguerite Ave, 381.54m away; The Fig Tree Restaurant, visit from 10:00 to 10:30, nearest transit: 7th St & Armory Dr, 103.61m away; Mint Museum Uptown, visit from 11:30 to 14:00, nearest transit: Tryon St & Stonewall St, 100.42m away; Princess Charlotte, visit from 14:30 to 15:30, nearest transit: Beaty St @ 866, 6465940.22m away; The Bank Of America Stadium, visit from 16:00 to 18:30, nearest transit: Stonewall St & Mint St, 270.52m away; Cabo Fish Taco Baja Seagrill, visit from 19:30 to 21:30, nearest transit: Davidson & 35th, 57.25m away; Tippah Treehouse Retreat, stay from 22:00 to 09:00, nearest transit: The Plaza & Marguerite Ave, 381.54m away."},
{"days": 5, "current_city": "from Charlotte to Lynchburg", ... }]}
Disruption Info:
Day 4 Attraction: The Bank of America Stadium
Disruption Category: Attractions
Reason: Stadium tours may be unavailable due to game-day prep or private events.
Disruption Detection Timestamp: 1 day before (via official stadium tour site)
Disruption Severity: Step-level
Confidence Level: Medium
Revised Plan {"acknowledgement": "Attractions Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.", "idx": 177, "row_number": "65-1", "JSON": {"org": "Lynchburg", "dest": "North Carolina", "days": 5, "visiting_city_number": 2, "date": ["2024-11-02", "2024-11-03", "2024-11-04", "2024-11-05", "2024-11-06"], "people_number": 1, "local_constraint": {"house rule": "parties", "cuisine": null, "room type": "entire home", "transportation": null, "event": null, "attraction": "Nature & Parks"}, "budget": 1800.0, "query": null, "level": "hard"}, "persona": "Traveler Type: Laidback Traveler; Purpose of Travel: Nature; Spending Preference: Economical Traveler; Location Preference: Beaches", "disruption tolerance": "Flexiventurer", "plan": [{"days": 1, "current_city": "from Lynchburg to Greensboro", ...}

{"acknowledgement": "Attractions Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.", "idx": 27, "row_number": "532-1", "JSON": {"org": "Miami", "dest": "Baltimore", "days": 3, "visiting_city_number": 1, "date": ["2024-11-03", "2024-11-04", "2024-11-05"], "people_number": 2, "local_constraint": {"house rule": null, "cuisine": null, "room type": null, "transportation": null, "event": ["Arts & Theatre"], "attraction": null}, "budget": 1800.0, "query": null, "level": "medium"}, "persona": "Traveler Type: Adventure Seeker; Purpose of Travel: Adventure; Spending Preference: Luxury Traveler; Location Preference: Mountains", "disruption tolerance": "Planbound", "plan": [{"days": 1, "current_city": "from Miami to Baltimore", ... },
{"days": 2, "current_city": "Baltimore", ... },
{"days": 4, "current_city": "Charlotte", "transportation": "-", "breakfast": "The Fig Tree Restaurant, Charlotte", "attraction": "Mint Museum Uptown, Charlotte; The Bank Of America Stadium, Charlotte", "lunch": "Princess Charlotte, Charlotte", "dinner": "Cabo Fish Taco Baja Seagrill, Charlotte", "accommodation": "Tippah Treehouse Retreat, Charlotte", "event": "-", "point_of_interest_list": " Tippah Treehouse Retreat, stay from 09:00 to 09:30, nearest transit: The Plaza & Marguerite Ave, 381.54m away; The Fig Tree Restaurant, visit from 10:00 to 10:30, nearest transit: 7th St & Armory Dr, 103.61m away; Mint Museum Uptown, visit from 11:30 to 14:00, nearest transit:

Tryon St & Stonewall St, 100.42m away; Princess Charlotte, visit from 14:30 to 15:30, nearest transit: Beaty St @ 866, 6465940.22m away; The Bank Of America Stadium, visit from 16:00 to 18:30, nearest transit: Stonewall St & Mint St, 270.52m away; Cabo Fish Taco Baja Seagrill, visit from 19:30 to 21:30, nearest transit: Davidson & 35th, 57.25m away; Tippah Treehouse Retreat, stay from 22:00 to 09:00, nearest transit: The Plaza& Marguerite Ave, 381.54m away."}, {"days": 5, "current_city": "from Charlotte to Lynchburg", ... }]}

Analysis: Even though the model acknowledges the Day 4 Disruption, but while generating the revised plan it fails to mitigate the Disrupted Attraction from POI list. This shows the declining Responsiveness rate when the Travel Duration increases.

G.3 Example showing the GPT-4o's model struggles to capture the Disruption Severity and Traveler Tolerance Level

Initial Plan {"idx": 37, "row_number": 401, "JSON": {"org": "Miami", "dest": "Louisville", "days": 3, "visiting_city_number": 1, "date": ["2024-11-21", "2024-11-22", "2024-11-23"], "people_number": 6, "local_constraint": "house rule": null, "cuisine": ["American", "Asian"], "room type": null, "event": null, "attraction": null, "transportation": null, "budget": 8250.0, "query": null, "level": "medium"}, "persona": "Traveler Type: Laidback Traveler; Purpose of Travel: Cultural Exploration; Spending Preference: Economical Traveler; Location Preference: Cities","disruption tolerance": "Planbound", "plan": [{"days": 1, "current_city": "from Miami to Louisville", "transportation": "Flight Number: F1028442, from Miami to Louisville, Departure Time: 15:29, Arrival Time: 18:13", "breakfast": "-", "attraction": "-", "lunch": "-", "dinner": "The Joy Luck, Louisville", "accommodation": "HIghlands Modern Get Away, Louisville", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 19:00 to 21:00, nearest transit: Taylorsville @ Bardstown, 142.01m away;The Joy Luck, visit from 21:30 to 22:30, nearest transit: Bardstown @ Longest, 70.75m away;HIghlands Modern Get Away, stay from 23:00 to 08:00, nearest transit: Taylorsville @ Bardstown, 142.01m away."}, {"days": 2, "current_city": "Louisville", "transportation": "-", "breakfast": "Jack Fry's, Louisville","attraction": "Muhammad Ali Center,Louisville; Frazier History Museum, Louisville; Waverly Hills Sanatorium, Louisville", "lunch": "English Grill, Louisville", "dinner": "Guy Fieri's Smokehouse, Louisville", "accommodation": "HIghlands Modern Get Away, Louisville", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 08:00 to 09:00, nearest transit: Taylorsville @ Bardstown, 142.01m away;Jack Fry's, visit from 09:30 to 10:00, nearest transit: Baxter @ Highland, 14.49m away;Muhammad Ali Center, visit from 11:00 to 13:00, nearest transit: Main @ 7th, 185.29m away;English Grill, visit from 14:00 to 14:30, nearest transit: Broadway @ 4th, 59.77m away;Frazier History Museum, visit from 15:00 to 17:00, nearest transit: Market @ S, 183.51m away;Waverly Hills Sanatorium, visit from 18:00 to 20:00, nearest transit: Dixie @ Maryman, 499.50m away;Guy Fieri's Smokehouse, visit from 21:00 to 22:00, nearest transit: Muhammad Ali @ 4th, 95.01m away;HIghlands Modern Get Away, stay from 22:30 to 05:00, nearest transit: Taylorsville @ Bardstown, 142.01m away."}, {"days": 3, "current_city": "from Louisville to Miami", "transportation": "Flight Number: F1369775, from Louisville to Miami, Departure Time: 07:02, Arrival Time: 09:18", "breakfast": "Porch Kitchen & Bar, Louisville", "attraction": "-", "lunch": "-", "dinner": "-", "accommodation": "-", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 05:00 to 05:30, nearest transit: Taylorsville @ Bardstown, 142.01m away;Porch Kitchen & Bar, visit from 06:00 to 06:30, nearest transit: W. Jefferson @ 2nd, 40.98m away."}]}

Disruption Info

Day 2 Attraction: Muhammad Ali Center, Louisville

Disruption Category: Attractions, Miscellaneous

Reason: The Muhammad Ali Center is closed until further notice due to fire damage to its HVAC system, making it inaccessible to visitors. Unexpected closures can also disrupt plans for cultural exploration and cause logistical stress

Disruption Detection Timestamp: 1 day before or upon arrival (via center website or local news)

Disruption Severity: Step-level

Confidence Level: High

Revised Plan

{"acknowledgement": "Attractions Disruption identified and understood. Proceeding with appropriate revisions based on disruption severity and traveler tolerance.", "idx": 37, "row_number": 401, "JSON": {"org": "Miami", "dest": "Louisville", "days": 3, "visiting_city_number": 1, "date": ["2024-11-21", "2024-11-22", "2024-11-23"], "people_number": 6, "local_constraint": {"house rule": null, "cuisine": ["American", "Asian"], "room type": null, "event": null, "attraction": null, "transportation": null }, "budget": 8250.0, "query": null, "level": "medium"}, "persona": "Traveler Type: Laidback Traveler; Purpose of Travel: Cultural Exploration; Spending Preference: Economical Traveler; Location Preference: Cities", "disruption_tolerance": "Planbound", "plan": [{"days": 1, "current_city": "from Miami to Louisville", "transportation": "Flight Number: F1028442, from Miami to Louisville, Departure Time: 15:29, Arrival Time: 18:13", "breakfast": "-", "attraction": "-", "lunch": "-", "dinner": "Eiffel Tower Restaurant, Louisville", "accommodation": "HIghlands Modern Get Away, Louisville", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 19:00 to 21:00, nearest transit: Taylorsville @ Bardstown, 142.01m away;Eiffel Tower Restaurant, visit from 21:30 to 22:30, nearest transit: The Strat, 4118.11m away;HIghlands Modern Get Away, stay from 23:00 to 08:00, nearest transit: Taylorsville @ Bardstown, 142.01m away."}, {"days": 2, "current_city": "Louisville", "transportation": "-", "breakfast": "Jack Fry's, Louisville", "attraction": "Frazier History Museum, Louisville; Waverly Hills Sanatorium, Louisville; Speed Art Museum, Louisville", "lunch": "English Grill, Louisville", "dinner": "Guy Fieri's Smokehouse, Louisville", "accommodation": "HIghlands Modern Get Away, Louisville", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 08:00 to 09:00, nearest transit: Taylorsville @ Bardstown, 142.01m away;Jack Fry's, visit from 09:30 to 10:00, nearest transit: Baxter @ Highland, 14.49m away;Frazier History Museum, visit from 11:00 to 13:00, nearest transit: Market @ S, 183.51m away;English Grill, visit from 14:00 to 14:30, nearest transit: Broadway @ 4th, 59.77m away;Speed Art Museum, visit from 15:00 to 17:00, nearest transit: S. 3rd @ Cardinal, 120.45m away;Waverly Hills Sanatorium, visit from 18:00 to 20:00, nearest transit: Dixie @ Maryman, 499.50m away;Guy Fieri's Smokehouse, visit from 21:00 to 22:00, nearest transit: Muhammad Ali @ 4th, 95.01m away;HIghlands Modern Get Away, stay from 22:30 to 05:00, nearest transit: Taylorsville @ Bardstown, 142.01m away."}, {"days": 3, "current_city": "from Louisville to Miami", "transportation": "Flight Number: F1369775, from Louisville to Miami, Departure Time: 07:02, Arrival Time: 09:18", "breakfast": "Porch Kitchen & Bar, Louisville", "attraction": "-", "lunch": "-", "dinner": "-", "accommodation": "-", "event": "-", "point_of_interest_list": "HIghlands Modern Get Away, stay from 05:00 to 05:30, nearest transit: Taylorsville @ Bardstown, 142.01m away;Porch Kitchen & Bar, visit from 06:00 to 06:30, nearest transit: W. Jefferson @ 2nd, 40.98m away."}] }

Analysis: We can observe the traveler has tolerance level of "Plan-Bound"with disruption severity of Step-level, but the LLM hallucinates and struggles to identify the relation and generates a revised plan which updates multiple days POI list.