TripTide: A Benchmark for Adaptive Travel Planning under Disruptions

Priyanshu Karmakar School of Electrical and Computer Sciences, IIT Bhubaneswar Bhubaneswar, India <a25cs09001@iitbbs.ac.in>

Manish Gupta
Microsoft
India
<gmanish@microsoft.com>

Soumyabrata Chaudhuri*
School of Electrical and Computer
Sciences, IIT Bhubaneswar
Bhubaneswar, India
<chaudhurisoumyabrata@gmail.com>

Abhik Jana
School of Electrical and Computer
Sciences, IIT Bhubaneswar
Bhubaneswar, India
<abhikjana@iitbbs.ac.in>

Shubhojit Mallick
Microsoft
India
<shmallic@microsoft.com>

Shreya Ghosh
School of Electrical and Computer
Sciences, IIT Bhubaneswar
Bhubaneswar, India
<shreya.cst@gmail.com>

Abstract

Recent efforts like TripCraft and TravelPlanner have advanced the use of Large Language Models (LLMs) for personalized, constraint-aware travel itinerary generation. However, real-world travel rarely unfolds without disruptions. To address this critical gap, we introduce TripTide¹, the first benchmark specifically designed to evaluate LLMs' ability to adapt itineraries in the face of realistic disruptions. TripTide models key dimensions such as disruption severity levels and traveler tolerance profiles, enabling nuanced assessment of LLM responses to unexpected events like transit cancellations, weather-related closures, or overbooked attractions.

We perform a holistic three-fold evaluation. First, we propose, a suite of novel automatic evaluation metrics, including 'Preservation of Intent' (assessing the feasibility and goal retention of revised plans), 'Responsiveness' (capturing how promptly and appropriately the LLM addresses disruptions), and a set of 'Adaptability' metrics that quantify semantic, spatial, and sequential divergence between original and modified itineraries. Second, we use LLM-as-a-judge approach to evaluate TripTide's effectiveness in handling disrupted plans. Third, a manual evaluation is conducted to verify whether the revision to address disruptions adequately captures the semantic, spatial, sequential, and responsiveness aspects.

Our experiments reveal intriguing mitigation strategies. LLMs tend to maintain strong sequentiality and stable semantic divergence between the original and revised plans, while the spatial deviations are pronounced for shorter plans but diminish for longer ones, suggesting that extended plan length allows better geographic coherence.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXXXXXXXXXXX

TripTide also reveals that the LLMs' capability to mitigate the disruptions gradually declines as plan horizons increase. TripTide² establishes a new foundation for benchmarking real-world robustness in travel planning agents, emphasizing adaptability, personalization, and resilience in LLM-generated itineraries.

ACM Reference Format:

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, existing works like *TravelPlanner* [24] and *TripCraft* [3] often assume an ideal, disruption-free environment. Other works, such as *Flex TravelPlanner* [12], have attempted to shed light on the topic, but have failed to capture the real-life disruption scenarios that occur during actual 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, closure of a point of interest, accommodation constraints, etc.), offering insights into their real-world utility and ability to revise a trip due to disruptions.

To fill this critical gap, we augment one of the existing datasets, namely TripCraft [3], 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

^{*}Currently at UT Austin, USA.

¹Capturing the ebb and flow of disrupted travel plans

²Codebase: https://anonymous.4open.science/r/TripTide-C3A7/

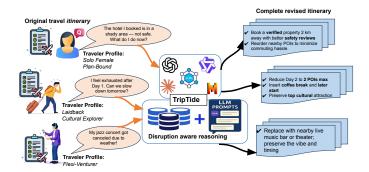


Figure 1: Motivating examples of persona-guided itinerary adaptation in TripTide under different types of disruptions

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 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.

	Subtype	Example / Description
	Transport	Flight Delays, cancellations, missed connections, baggage issues, strikes, weather, or airport closures
_	Accommodation	Poor hygiene, unsafe/remote location, missing promised amenities
legory	Restaurant	Closures, changes in hours, poor hygiene, no dietary support, overcrowding, long waits, surge pricing
Disruption Category	Attractions	Closed for maintenance, weather-related issues, government restrictions, ticket sell-outs, ID rules, overcrowding, limited open days.
Disrup	Miscellaneous	Any event based disruptions like cancellation of events, rescheduling a event, sudden venue changes, event ticket issues, weather or safety risks, also includes any other form of disruptions that has not been categorized like illness, jet lag, fatigue, bad weather, forgotten documents.
Severity	Step-Level Day-Level Plan-Level	Affects an individual activity or segment Impacts all activities on a particular day Necessitates major plan or route overhauls
Traveler Severity Toler- ance	Flexi-Venturer Plan-Bound	Open to rerouting, substitution, or plan changes Prefers minimal changes and high plan adherence following Disruption severity.

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

(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.

Collectively, these contributions lay the foundation for a new standard in evaluating and improving LLM-driven travel planning systems, particularly in the context of real-world scenarios where disruptions are the norm.

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 [1, 20]. LLMs have shown impressive capabilities in task decomposition, commonsense reasoning, and step-by-step decision-making [21, 25], making them compelling candidates for real-world planning applications. Recent approaches such as LLM-Planner [17] enable few-shot grounded planning by iteratively refining high-level plans based on feedback. At the same time, hybrid methods integrating LLMs with classical algorithms like Monte Carlo Tree Search have demonstrated improved planning efficiency [19, 27]. Other existing works like TripTailor [15] and [14] 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 openended environments [26], often faltering in sub-goal coordination and long-horizon dependencies [10, 16, 20]. While chain-of-thought prompting [21] and fine-tuning improve benchmark performance, they falter on novel, complex tasks. Retrieval-based methods such as RAG [11] and knowledge graphs [18, 23] likewise remain unproven in real-world travel disruptions scenarios.

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 [7]. The recent application of LLMs to this domain has sparked interest due to their ability to interpret and generate natural language instructions [9, 22]. Initial benchmarks, such as the one proposed by Xie et al. [24], highlight the difficulty LLMs face in balancing

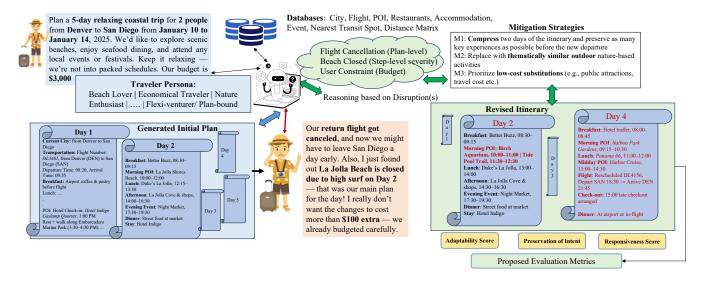


Figure 2: Illustration of multi-level disruption handling and adaptive travel replanning by TripTide

multiple constraints [2], often producing itineraries that violate feasibility or preference criteria. Although subsequent works [6, 8] report promising results, they typically oversimplify the problem by omitting real-world limitations such as transit schedules, event timing, and dynamic user contexts [13] and using discrete binary evaluation metrics. Although recent efforts [3, 4, 16] 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 [3], we generate three potential disruptions per

day of the itinerary by prompting GPT-4o and Gemini 2.5 Pro [5]. Each disruption is then paired with a corresponding disruption query, which is used as input to prompt GPT-4o³ to generate a persona-aware revised plan. The complete prompts are mentioned in our codebase². Hyperparameter details are provided in Appendix A.

Disruption Categories and Scenarios. This section outlines the various categories of disruptions that may occur during a trip, along with representative scenarios under each category. These are also illustrated in Table 1.

- **1. Flight Disruptions.** These include flight delays, cancellations, missed connections, baggage mishandling, labour strikes, adverse weather conditions, and airport closures.
- **2. Accommodation Disruptions.** These can arise due to poor hygiene, unsafe or remote locations, unavailability of promised amenities, or overbooking.
- **3. Restaurant Disruptions.** These include unexpected closures, changes in operating hours, poor hygiene standards, inadequate dietary accommodations, overcrowding, lengthy wait times, or surge pricing.
- **4. Event and Activity Disruptions.** These include event cancellations, rescheduling, venue changes, ticketing issues, or safety-related concerns due to weather or local advisories.
- **5. Miscellaneous Disruptions.** Other disruptions encompass traveler illness, jet lag, fatigue, inclement weather, or forgotten documents, all of which may affect the overall itinerary.

Disruption Severity Levels. As a key feature of 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

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

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: the Flexi-Venturer, who readily adapts to changes, and the *Plan-Bound* traveler, who 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.

Flexi-Venturer Traveler. This traveler exhibits high adaptability and is open to changes in the itinerary following a disruption, regardless of its scope. They are comfortable with dynamic re-planning and show low sensitivity to deviations from the original plan.

Plan-Bound Traveler. In contrast, the Plan-Bound traveler is rigid and prefers to adhere strictly to the original itinerary. Their acceptance of changes is constrained by the disruption severity. For instance, in the case of a step-level disruption, they permit only that specific step to be modified.

3.2 Annotation Construction 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 annotators⁴ 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-40 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 B.

Constraint Dimension	Subtype	Pl	Total		
		3-day	5-day	7-day	
	Transport	118	364	555	1,037
	Accomodation	225	416	418	1,059
D G.	Restaurant	1,072	1,223	1,618	3,913
Disruption Category	Attractions	980	1,223	1,618	3,821
	Miscellaneous	94	437	697	1,228
	Total	2,489	3,663	4,906	11,058
	Step-level	111	140	134	385
D: 41 G 14	Day-level	117	111	131	359
Disruption Severity	Plan-level	116	73	67	256
	Total	344	324	332	1,000
	Flexi-Venturer	172	162	166	500
Traveler Tolerance	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., 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

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

1. Preservation of Intent. This metric evaluates whether the user's original travel intent, as characterized by constraints and preferences,

⁴Undergraduate interns at our lab

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 [24] 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:

Responsiveness Rate =
$$\frac{\text{#Mitigated Plans}}{\text{#Total Plans}}$$
 (1)

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, each measured as follows:
- i) Semantic Adaptability ($A_{\rm sem}$). 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 [3], in our setting, it is repurposed to reflect plan-level thematic continuity rather than alignment with the user persona:

$$\bar{S}_{ps} = \frac{1}{M \cdot N} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\mathbf{p}_{j} \cdot \mathbf{q}_{i}}{\|\mathbf{p}_{j}\| \|\mathbf{q}_{i}\|}$$
 (2)

where p_j and 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 then:

$$A_{\text{sem}} = \left| \bar{S}_{\text{ps}}^{\text{i}} - \bar{S}_{\text{ps}}^{\text{r}} \right| \tag{3}$$

where \bar{S}^{i}_{ps} and \bar{S}^{r}_{ps} are the persona scores of the initial and revised plans, respectively. Lower values are preferred.

ii) Spatial Adaptability ($A_{\rm spa}$). This measures the difference in spatial convenience between the original and revised plans. Following [3], we compute the spatial score for each plan using:

$$S_s(d) = \begin{cases} 1 - 0.5 \left(\frac{d}{d_0}\right), & \text{if } d \le d_0 \\ 0.5 \exp\left(-\lambda(d - d_0)\right), & \text{if } d > d_0 \end{cases}$$
 (4)

$$\bar{S}_{\text{spatial}} = \frac{1}{N} \sum_{i=1}^{N} S_{s}(d_{i})$$
 (5)

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:

$$A_{\rm spa} = \left| \bar{S}_{\rm spatial}^{\rm i} - \bar{S}_{\rm spatial}^{\rm r} \right| \tag{6}$$

where $\bar{S}^{i}_{spatial}$ and $\bar{S}^{r}_{spatial}$ are the spatial scores of the initial and revised plans, respectively. Lower values are preferred.

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

$$A_{\text{seq}} = \frac{\text{ED}(\mathcal{G}, \mathcal{A})}{\max(|\mathcal{G}|, |\mathcal{A}|)}$$
(7)

where \mathcal{G} and \mathcal{A} are PoI sequences of the initial and revised plans, and $\mathrm{ED}(\cdot,\cdot)$ is the edit distance. We average this over all days of both the plans to get A_{seq} . Lower values are preferred.

4.2 Results

We report the performance⁵ of GPT-40 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-40 and Qwen2.5-7B-Instruct achieve high delivery rates (≥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-40 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→18.89→7.67, macro 24.49→13.28→7.96). This suggests that Qwen2.5-7B-Instruct and Phi4-mini-Instruct's constraint reasoning ability diminishes with increasing plan complexity, while GPT-40 maintains more stable performance. In terms of final pass rates, GPT-40 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 ~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-4mini Instruct's reliability in producing plans, while also underscoring the need for improved mechanisms to ensure comprehensive constraint satisfaction.

Semantic Adaptability ($A_{\rm sem}$). GPT-40 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-40. 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

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

Model Name	Plan Duration	Delivery Rate (%)↑	CPR (%)↑		HCPR (%)↑		Final Pass	Adaptability (%)			Responsive-
			Micro	Macro	Micro	Macro	Rate (%)↑	Semantic Score $(A_{\text{sem}}) \downarrow$	Spatial Score $(A_{\mathrm{spa}}) \downarrow$	$\begin{array}{c} \textbf{Sequential Score} \\ (A_{\text{seq}}) \downarrow \end{array}$	ness Rate (%)↑
	3-day	100.00	91.02	29.95	50.93	48.84	29.37	0.34	1.22	10.89	89.53
GPT-40	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
	3-day	100.00	89.56	31.16	54.02	51.93	29.98	0.67	5.94	56.26	97.67
Qwen2.5-7B-Instruct	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
	3-day	45.23	39.75	18.68	20.36	24.49	18.26	26.27	3.32	26.64	75.94
Phi-4 mini Instruct	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 percentage points.

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*_{spa}). For GPT-40, 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-40 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 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-40. 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_{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-40 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-40 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-40 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-40 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-40, 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.

Summary. GPT-40 demonstrates robust semantic preservation and modest spatial reorganization capacity. Its ability to mitigate disruptions decreases slightly with longer durations, and its static sequential performance reflects a need for stronger modeling of temporal dependencies. Qwen2.5-7B-Instruct demonstrates stronger responsiveness in longer itineraries than GPT-40. Phi-4-mini often fails to deliver accurate plans when disruptions occur, but it correctly identifies the disruption. Additional result analysis is in Appendix C.

5 LLM-as-a-Judge based Evaluation

In this section, we present an orthogonal, Llama-as-judge perspective on GPT-40 revised plans, summarizing perceived quality across 3/5/7-day itineraries.

5.1 Experimental Setup

In addition to the novel evaluation metrics discussed in Section 4, we further analyze the generated travel itineraries using a separate LLM. Motivated by the idea of *LLM-as-a-Judge* [28], we adapt our prompt and evaluate the responses accordingly. As shown in Table 3, GPT-40 consistently achieves the best average performance across our evaluation metrics. Hence, we perform this evaluation against GPT-40 outputs.

For each user query, we provide the original plan, the disruption information, and the revised plan generated by the model. We then rate the quality of the revised plan on a scale of 1 to 5, where a score of 1 indicates the worst case (such as when the regenerated plan is nearly identical to the original or fails to meet the user

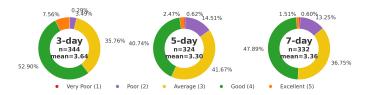


Figure 3: Llama-3.1 8B Instruct Evaluation Results

persona's needs) while 5 indicates the best case (disruption is fixed fully). The detailed scoring scheme is presented in Table 4. We perform this evaluation using the Llama-3.1-8B-Instruct⁶ LLM. By evaluating GPT-4o's responses in this manner, and by applying a carefully designed evaluation prompt, we were able to reduce bias and minimize hallucination in the assessment process.

5.2 Results

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 *AveragelGood* 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 ($\approx 100\%$), 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 Rate 41.98%. 3-day plans, while best judged (mean 3.64), show weaker formal validity: CPR_{macro} of 29.95, HCPR_{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

To complement automated scoring, we conducted a human evaluation to assess LLM efficiency within our disruption-aware framework. Three domain experts independently reviewed 100 revised itineraries (sampled across 3-, 5-, and 7-day plans). For each case, they examined the original plan, the disruption, and the revision, judging responsiveness and whether updates preserved semantic, spatial, sequential aspects.

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

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 4: Rubric for LLM judging of disrupted travel plans

```
Traveler Persona: {"type": "Laidback", 'people number": "7", "budget": "Luxury Traveler", "Tolerance": "Planbound"}
Original Plan
{"day": 1, "city": "Los Angeles → Colorado", } {"day": 2, ... } {"day": 3, ... } ...
Disruption: Miscellaneous Ground Transportation Disruption
{"Day 1 Transport", "Reason": "The original plan does not support Transportation for seven people."}
Revised Plan
{"day": 1, "transport": "Pre-booked large passenger van for airport transfers and city travel in Colorado Springs."} {"day": 2, ... } {"day": 3, ... } ...
Insight: The Planner proactively added the missing transport on Day 1 in revised plan.
```

Figure 4: Enhanced plan, handling logistics for a large group.

```
Traveler Persona: {"type": "Laidback", "location": "Nature", "Tolerance": "Flexiventurer' }
Original Plan
{"day": 1, "city": "Boise → California", }, {"day": 2, ... }, {"day": 3, "attraction": "Forestiere
Underground Gardens, Fresno, at 18:00 pm" } ...
Disruption: Attraction Disruption
{"Day 3 Attraction", "Reason": "Tour timing infeasible due to darkness."}
Revised Plan
{"day": 1, ...}, {"day": 2, ... } {"day": 3, "attraction": Fresno Chaffee Zoo, Fresno, at 19:30pm" } ...
Insight: The Planner replaces a late-night attraction to a zoo at night.
```

Figure 5: Planner misses suitable alternative.

The study focused on how models like GPT-40, Qwen-2.5 7B Instruct, and Phi-4-mini Instruct demonstrate reasoning, contextual understanding, and adaptability under unexpected changes. These capabilities enable dynamic, user-aligned revisions that account for real-world uncertainties, setting a new benchmark for disruption-aware trip generation. Expert judgments form the basis of the analysis reported in this section. Detailed case studies are in Appendix D.

6.1 General trends

Reliable disruption detection, disciplined yet non-trivial revisions. Across all samples, the GPT-40 planner consistently identified the disrupted element (venue closure, transport capacity, infeasible timing) and initiated a corrective edit. Experts noted that, 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 frequently 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 the 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 (see Fig. 4).

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 undercorrection 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. See Fig. 5 for a timing miss where the alternative remained infeasible.

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. A consistent mitigation pattern was observed for accommodation-related disruptions. In four out of five instances, the unsuitable accommodation was directly replaced with a more appropriate alternative, tailored to traveler preferences such as size, comfort, or luxury level. In one case, instead of replacing the accommodation, the subsequent day's itinerary was adapted to accommodate traveler fatigue from a late-night check-in. These strategies reflect a high degree of responsiveness, as the system proactively addressed the disruption. However, slight reductions in the sequentiality score were observed when overnight adjustments resulted in deviations from the following day's schedule.

Transportation Disruptions. Transportation issues were typically resolved with minimal disruption to the overall itinerary structure. Adjustments included modifying departure times, inserting rest breaks, or switching to more appropriate transportation modes (e.g., from public to self-driven). These interventions preserved both the feasibility and logical flow of travel, contributing to consistently

```
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
```

Figure 6: Semantic issues for a 5-day plan

Mexican style restaurant by a French type restaurant

```
Traveler Persona: {"type": "Laidback", "purpose": "History", "budget": "Economical", "pref "Museums", "Tolerance": "Plan-Bound"}
 {"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
Spatial Divergence 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 divergence A_{SPA}
```

Figure 7: Preservation of spatial convenience for a 3 day plan

high semantic alignment and spatial coherence. While such decisions showcased a strong contextual understanding, the model occasionally opted for conservative changes, resulting in slightly lower responsiveness in cases of 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. Adjustments were also made to better align with traveler profiles, for instance, replacing luxury accommodations for budget travelers, and vice versa. These nuanced modifications resulted in elevated semantic scores, reflecting the model's sensitivity to user-specific needs. However, such refinements occasionally

came at the expense of spatial efficiency, particularly when suitable alternatives were not in close proximity.

Summary. Overall, the analysis demonstrates that the model performs well across most disruption types, exhibiting particularly strong performance in responsiveness and semantic preservation. Minor trade-offs in sequentiality and spatial planning were observed in specific scenarios, often arising from the inherent complexity of balancing user intent with contextual feasibility. These findings highlight the model's ability to effectively adapt plans in a user-aligned and disruption-aware manner.

Analysis of Examples. The two snippets in Figs. 6 and 7 show the robust performance of GPT-40 for a 5-day and a 3-day case from our dataset. The model makes multiple semantic mistakes in 5-day while maintaining the spatial coherence for 3-day. This snippet indicates substantial scope for further improvement in this area.

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.

References

- [1] 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. arXiv:2406.13094 [cs.CL] https://arxiv.org/abs/2406.13094
- [2] Andrey Borro, Patricia J Riddle, Michael W Barley, and Michael J Witbrock. 2025.
 Large Language Models as Common-Sense Heuristics. arXiv:2501.18816 [cs.CL] https://arxiv.org/abs/2501.18816
- [3] 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). Association for Computational Linguistics, Vienna, Austria, 17035– 17064. https://aclanthology.org/2025.acl-long.834/
- [4] Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. 2024. TravelAgent: An AI Assistant for Personalized Travel Planning. arXiv:2409.08069 [cs.AI] https://arxiv.org/abs/2409.08069
- [5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025).
- [6] 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 (2024).
- [7] Omkar Gurjar and Manish Gupta. 2021. Should I visit this place? Inclusion and exclusion phrase mining from reviews. In European Conference on Information Retrieval. Springer, 287–294.

- [8] Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. 2025. Large Language Models Can Solve Real-World Planning Rigorously with Formal Verification Tools. arXiv:2404.11891 [cs.AI] https://arxiv.org/abs/2404.11891
- [9] 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 (2025).
- [10] Subbarao. Kambhampati, Karthik. Valmeekam, Matthew. Marquez, and Lin. Guan. 2023. On the Role of Large Language Models in Planning. https://yochan-lab. github.io/tutorial/ICAPS-2023/ Tutorial presented at the International Conference on Automated Planning and Scheduling (ICAPS), Prague.
- [11] 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. arXiv:2504.08694 [cs.CL] https://arxiv.org/abs/2504.08694
- [12] 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. https://openreview.net/forum?id=a7unQ5jMx7
- [13] 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 (2024).
- [14] 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. arXiv:2412.13682 [cs.AI] https://arxiv.org/abs/2412.13682
- [15] Yuanzhe Shen, Kaimin Wang, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2025. TripTailor: A Real-World Benchmark for Personalized Travel Planning. arXiv:2508.01432 [cs.AI] https://arxiv.org/abs/2508.01432
- [16] 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, Franck Dernoncourt, Daniel Preotiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 486–514. doi:10.18653/v1/2024.emnlp-industry.37
- [17] 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)*.
- [18] 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 (2024). doi:10.3390/ijgi13110414
- [19] 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 (2023), 2497–2562.
- [20] 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). arXiv:2302.06706 [cs.AI] https://arxiv.org/abs/2302.06706
- [21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35 (2022), 24824–24837.
- [22] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. Science China Information Sciences 68, 2 (2025), 121101.
- [23] 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, Maosong Sun, Sujian Li, Yue Zhang, and Yang Liu (Eds.). Chinese Information Processing Society of China, Haikou, China, 939–949. https://aclanthology.org/2020.ccl-1.87/
- [24] 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.
- [25] 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*. Springer, 24–38.
- [26] 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, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE,

10054-10082. https://aclanthology.org/2025.coling-main.672/

- [27] 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 (2024).
- [28] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging Ilm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems 36 (2023), 46595–46623.

Overview of Appendix Sections

- Appendix A: Hyper-parameter Details for Reproducibility
- Appendix B: Human Annotation Guidelines
- Appendix C: Detailed Analysis of Results
- Appendix D: Case Studies

A 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-40	Qwen2.5-7B instruct	Phi-4- mini instruct
Model Source	OpenAI	Alibaba (via Hugging-	Microsoft (via Hugging-
		Face)	Face)
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 5: 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: 3.10Inference Libraries:
- transformers==4.53.2
- torch==2.5.1
- accelerate==1.9.0
- Memory: 46 GB GPU memory, 1TiB system RAM
- Model Hosting:
 - GPT-40: 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 / Plan-
	bound) when modifying the plan, and should provide clear remarks explaining the adjust-
	ments 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
0	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
10	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 6: Guidelines for Annotation of Revised Travel Plans and Remarks

B 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 6 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.

C Detailed 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 7 and 8.

Preservation of Intent: Table 7 shows that GPT-40 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-40 consistently achieves better constraint satisfaction overall indicating the model's strength to capture the intent across all the disruption severity levels.

Semantic Adaptability (A_{sem}): Table 8 reflects that GPT-40 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_{spa}): GPT-40 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.

Sequential Adaptability (A_{seq}): GPT-40 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-40 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-40 while checking the mitigation rate of the models between the final revised and initial plan. We can observe from Table 8 that for Qwen2.5-7B-Instruct the rate is nearly 100% while GPT-40 performance degrades as the severity level increases.

Summary. In this paper we analyzed capacity of GPT-40, 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 7 and Table 8, we can infer that, in the day-wise setting, GPT-40 maintains high sequential consistency and improving spatial scores with longer plans, showing stable adaptability. Qwen2.5-7B-Instruct and Phi-4-mini-Instruct

Model	Plan	Delivery	CPR↑		HCPR↑		Final Pass
Name	Duration	Rate↑	Micro	Macro	Micro	Macro	Rate↑
	Step-Level	100.00	90.29	32.46	47.68	46.56	31.15
GPT-40	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
Owen2.5-	Step Level	99.65	86.98	24.49	42.47	44.22	24.49
7B-Instruct	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 7: Preservation of user intent scores for different Models based on Disruption Severity level.

Model Name	Disruption Severity	Semantic Score $(A_{sem}) \downarrow$	Spatial Score $(A_{\operatorname{Spa}}) \downarrow$		Responsive- ness Rate ↑
	Step-level	0.05	0.65	82.36	94.11
GPT-40	Day-level	0.41	0.21	4.98	78.84
	Plan-level	0.17	1.41	4.55	76.76
Qwen2.5-7B-	Step-level	0.40	5.49	34.06	96.39
-	Day-level	1.46	0.01	32.54	99.67
instruct	Plan-level	0.41	1.31	31.42	99.57

Table 8: Adaptability $(A_{sem}, A_{spa}, A_{seq})$ and Responsiveness scores (%) for different models based on Disruption Severity level.

achieve higher responsiveness but struggle with semantic and spatial coherence, especially for 3-day plans. In the severity-wise setting, GPT-40 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.

D Case Studies

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

D.1 Example showing how GPT-40 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 Mon-ument 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 Monu 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

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 Daylevel disruption with another apprpriate attraction type. The model was also able to capture the tolerance level of traveler (Plan-bound) as all other day's POI list remains same.

D.2 Example citing the GPT-40 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", "2024-11-06"], "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": "Elexiventurer": "plan": "("dayse": 1 "current_city"). 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 Treeho Retreat, stay from 22:00 to 09:00, nearest transit: The Plaza & Marguerite Ave, 381.54m

{"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": "Nor Carolina", "days": 5, "visiting_city_number": 2, "date": ["2024-11-02", "2024-11-03", "2024-11-04", "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"], "peo-ple_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 Treeh Retreat, stay from 22:00 to 09:00, nearest transit: The Plaza& Marguerite Ave, 381.54m

{"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.

Example showing the GPT-40's model struggles to capture the Disruption Severity and Traveler Tolerance Level

Initial Plan {"idx": 37, "row_number": 401, "JSON": {"org": "Miami", " "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 Prefer-Type: Latoback Traveler; Purpose of Travel: Cultural Exploration; Spending Freference: 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": ".", "ulnuch": ".", "dinner": "The Joy Luck, Louisville", "accommodation": "Hghlands 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, 142.01 m away."],
["days": 2, "current_city": "Louisville", "transportation": "-", "breakfast": "Jack Fry's,
Louisville", "attraction": "Muhammad Ali Center,Louisville; Frazier History Museum, Louisville", "attraction": "Muhammad Alt 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

Attraction: Muhammad Ali Center, Louisville

Day 2 Attaction Audition 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

Confidence Level: High

Revised Plan

{"acknowledgement": "Attractions Disruption identified and understood. Proceeding acknowledgemine: Attractions Distuption 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: [Gays 1], Culrent_city: I non-Whalmi to Louisville, Departure Time: 15:29, Arrival Time: 18:13", "breakfast": "-", "attraction": "-", "lunch": "-", "dinner": "Eiffel Tower Restaurant, Louisville", "accommodation": "Hlghlands Modern Get Away, Louisville", "event": "-", "point_of_interest_list": "Hlghlands 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 -, accommodation: -, event: -, point_ol_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.