

TripCraft: A Benchmark for Spatio-Temporally Fine Grained Travel Planning

Soumyabrata Chaudhuri¹, Pranav Purkar¹, Ritwik Raghav¹, Shubhojit Mallick²

Manish Gupta², Abhik Jana¹, Shreya Ghosh¹

¹IIT Bhubaneswar, India ²Microsoft, India

chaudhurisoumyabrata@gmail.com

{23cs06011, a23cs09001, abhikjana, shreya}@iitbbs.ac.in

{shubhojit.mallick, gmanish}@microsoft.com

Abstract

Recent studies on probing Large Language Models (LLMs) have explored their latent potential as personalized travel planning agents, yet existing benchmarks remain limited in real-world applicability. Existing datasets, such as *TravelPlanner* and *TravelPlanner+*, suffer from semi-synthetic data reliance, spatial inconsistencies, and a lack of key travel constraints, making them inadequate for practical itinerary generation. To address these gaps, we introduce **TripCraft**, a spatio-temporally coherent travel planning dataset that integrates real-world constraints, including public transit schedules, event availability, diverse attraction categories, and user personas for enhanced personalization. To evaluate LLM-generated plans beyond existing binary validation methods, we propose five continuous evaluation metrics, namely Temporal Meal Score, Temporal Attraction Score, Spatial Score, Ordering Score, and Persona Score—which assess itinerary quality across multiple dimensions. Our parameter-informed setting significantly enhances meal scheduling, improving the Temporal Meal Score from 61% to 80% in a 7-day scenario. *TripCraft*¹ establishes a new benchmark for LLM-driven personalized travel planning, offering a more realistic, constraint-aware framework for itinerary generation.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in structured reasoning and decision-making (Wei et al., 2022; Yao et al., 2023), making them promising candidates for automated planning tasks. Recently, LLMs have been explored to generate travel itineraries, where the goal is to create coherent, personalized, and logistically sound travel plans. However, existing benchmarks such as *TravelPlanner* (Xie et al.,

2024) and *TravelPlanner+* (Singh et al., 2024) suffer from several limitations that restrict their real-world applicability. First, these datasets rely heavily on semi-synthetic data, leading to spatial inconsistencies. For instance, *TravelPlanner* assigns accommodations in New York across 312 different cities, causing unrealistic itinerary generation. Second, key real-world travel constraints, such as public transit schedules, event availability, attraction categories, and user preferences—are either missing or insufficiently modeled (Xie et al., 2024; Chen et al., 2024; Singh et al., 2024). While *TravelPlanner+* incorporates user personas, it lacks detailed modeling of persona-driven attraction selection, travel pace (adventurous vs laidback), and scheduling preferences, limiting its ability to produce customized plans.

To address these limitations, we introduce *TripCraft*, a benchmark for spatio-temporally fine-grained travel planning. Unlike prior datasets, *TripCraft* is constructed entirely from real-world data sources, ensuring geographic consistency, valid transit connectivity, and contextually accurate event and attraction information. It integrates public transit stops and schedules, enabling realistic transit-aware itinerary generation, and incorporates diverse attraction and event categories, such as concerts, museums, and sports, to enhance itinerary richness. Additionally, *TripCraft* models user personas, capturing travel styles, budget preferences, and location affinities to enable a more personalized and contextually relevant trip planning experience. Our dataset consists of **1000 travel queries** spanning 140 U.S. cities, covering 3-day, 5-day, and 7-day itineraries. Gold-standard plans are annotated by 25 human annotators through multiple refinement rounds, with detailed remarks to enhance interpretability, acknowledging multiple valid itineraries per query.

Evaluating LLM-generated travel plans is challenging, as existing methods rely on binary

¹Dataset and code available at <https://github.com/Soumyabrata2003/TripCraft>.

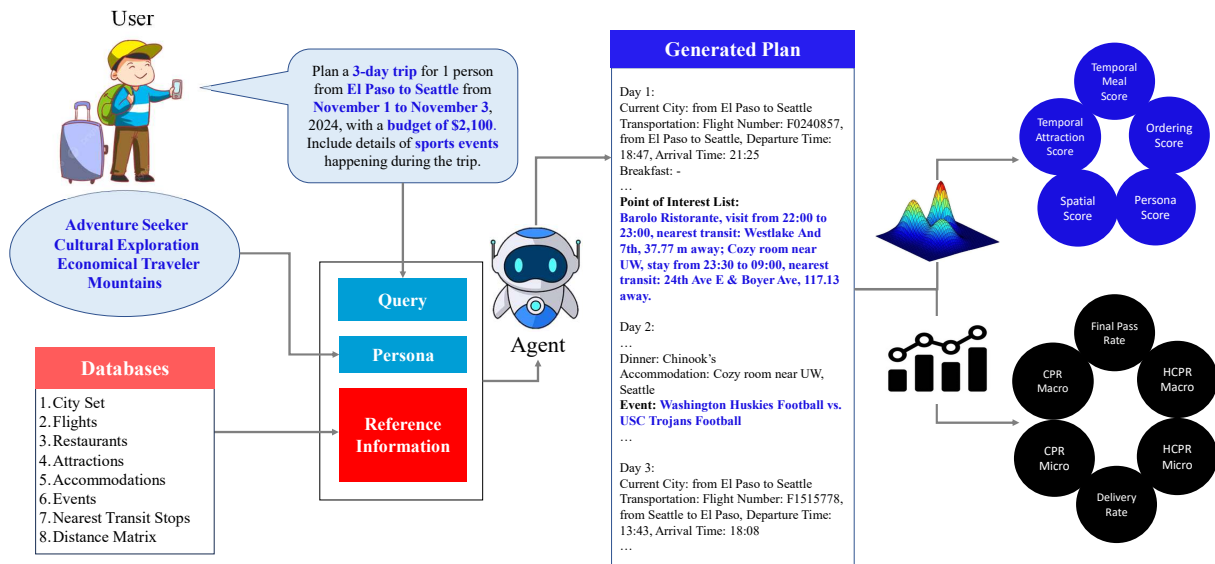


Figure 1: TripCraft overview: Continuous evaluation metrics along with fine-grained itinerary details.

checks—whether constraints are met or if one plan is better than the other—without capturing finer details. *But what truly makes a travel plan effective? Is it ensuring that meal schedules align with natural times? Does each attraction visit allow enough time for meaningful exploration, or is it unrealistically short or excessively long? Is travel between locations efficient, minimizing unnecessary detours and delays? Does the itinerary follow a logical sequence, or does it jump between distant places without consideration for practicality? Most importantly, does the plan align with personal preferences—catering to adventure seekers, cultural enthusiasts, or those looking for relaxation?* To systematically assess these aspects, we propose five novel evaluation metrics: *Temporal Meal Score*, ensuring natural meal scheduling; *Temporal Attraction Score*, evaluating visit durations; *Spatial Score*, measuring travel efficiency; *Ordering Score*, assessing logical sequencing; and *Persona Score*, capturing alignment with user preferences. These metrics move beyond binary validation (Xie et al., 2024; Singh et al., 2024), providing a continuous and explainable framework for assessing the quality of itineraries. Fig. 1 shows an overview of the proposed TripCraft benchmark. In a nutshell, our key contributions are three-fold:

1. **Fine-grained travel planning dataset:** TripCraft leverages real-world data to eliminate geographic inconsistencies, ensuring coherent plans. Our dataset incorporates detailed attraction types, persona-based preferences, event categories, and public

transit information, making it significantly richer than prior benchmarks.

2. **Highly detailed itinerary generation:** Our approach generates travel itineraries with enhanced spatio-temporal granularity, producing a comprehensive and contextually coherent sequence of Points of Interest (PoIs) that provides a holistic representation of the journey.
3. **Novel continuous evaluation metrics:** To the best of our knowledge, this is the first ever attempt to introduce continuous evaluation metrics for LLM-generated travel plans, moving beyond binary constraints to assess itinerary quality with greater nuance.

By addressing fundamental limitations in prior datasets and evaluation strategies, TripCraft sets a new standard for benchmarking LLMs in travel planning, enabling more detailed, explainable, and personalized itinerary generation.

2 Related Work

Planning with LLMs. LLMs have demonstrated significant potential in various planning tasks, including task scheduling, heuristic guidance, and commonsense reasoning (Borro et al., 2025; Huang et al., 2024; Valmeekam et al., 2023; Prasad et al., 2024; Pallagani et al., 2023; Lee et al., 2025). Song et al. (2023) introduced LLM-Planner which does few-shot grounded planning, dynamically updating high-level plans based on real-time feedback. Zhao et al. (2024) showed that integrating

Dimension	TravelPlanner	TravelPlanner+	TravelAgent	TripCraft (Ours)
Geographical coherence	✗	✗	✓	✓
Size	1225 queries	1225 queries	20 travel scenarios	1000 queries
Persona	✗	✓	✗	✓
Public transit data	✗	✗	✗	✓
Event data	✗	✗	✗	✓
Attraction categories	✗	✗	✗	✓
Constraints (C: Commonsense, H: Hard)	8C, 5H	8C, 5H	N/A	10C, 7H
PoI list in itinerary output	✗	✗	Temporal only	✓
Explanations/ Remarks	✗	✗	✗	✓
Evaluation metrics	Delivery Rate, CPR, HCPR, Final Pass Rate	Delivery Rate, CPR, HCPR, Final Pass Rate, Preference Rate (+)	Rationality, Comprehensiveness and Personalization	Delivery Rate, CPR, HCPR, Final Pass Rate, Temporal Meal Score (+), Temporal Attraction Score (+), Spatial Score (+), Persona Score (+), Ordering Score (+)

Table 1: Comparison of existing travel planning datasets across multiple qualitative and quantitative dimensions.

LLMs with classical planning techniques, such as Monte Carlo Tree Search (MCTS) (Coulom, 2006; Świechowski et al., 2023), enhances task-planning efficiency. However, despite their promise, LLMs struggle with generating effective plans independently across diverse domains (Valmeekam et al., 2023). Moreover, they face challenges in handling subgoal dependencies and require external reasoning mechanisms for robust planning (Kambhampati et al., 2023). Techniques such as chain-of-thought prompting and fine-tuning can improve performance but expose limitations when encountering novel, complex scenarios (Yang and Tomar, 2023; Bohnet et al., 2024).

LLMs in Travel Planning. Automated travel planning is inherently complex, requiring the optimization of multiple subgoals such as scheduling, budgeting, and route efficiency, while also incorporating user preferences (Gurjar and Gupta, 2021). The emergence of LLMs presents an opportunity to streamline this process through natural language interaction (Xi et al., 2025; Jonnala et al., 2025). Xie et al. (2024) introduced a benchmark with 1,225 travel-related queries, assessing LLMs against eight commonsense and five hard constraints. Their study revealed that LLMs struggle with multi-constraint optimization, leading to suboptimal travel plans. While papers like (Hao et al., 2025) and (Gundawar et al., 2024) reported strong performance in travel planning, their methodology omitted key constraints, simplifying the task. A major limitation in these studies is the absence of real-world datasets that incorporate public transit schedules, event calendars, and personalization factors, restricting their applicability

(Shao et al., 2024).

Evaluation of LLM-Generated Travel Plans. Existing evaluations of LLM-based travel planning rely on discrete constraint-checking methodologies. Metrics such as Delivery Rate measure an LLM’s ability to generate a plan without failure, while Xie et al. (2024) introduced assessments for explicit user requirements and implicit real-world feasibility. If a plan met all three criteria, it was deemed viable. Subsequent studies have built upon this framework; for instance, Chen et al. (2024) introduced metrics for rationality, personalization, and comprehensiveness, while Singh et al. (2024) proposed the Preference Rate metric, quantifying how often a personalized plan was favored over a generic one. However, evaluating travel plans solely based on constraint adherence is insufficient. A robust assessment must consider temporal, spatial, and sequential coherence—dimensions largely overlooked in prior works. Our proposed benchmark, TripCraft, addresses this gap by introducing continuous evaluation metrics that provide a fine-grained analysis of itinerary quality. Table 1 provides a comprehensive comparison between existing datasets and TripCraft, highlighting key differences and advantages.

3 TripCraft

3.1 Overview

We introduce TripCraft, a benchmark for evaluating language agents in complex, constraint-aware planning, specifically in travel itinerary generation. TripCraft assesses agents’ ability to construct flexible yet constrained plans by incorporating user preferences, commonsense considera-

tions, and persona-driven requirements while ensuring temporal, spatial, sequential, and personal consistency. The benchmark comprises 1000 diverse travel queries, categorized into three trip durations (3-day, 5-day, and 7-day) and three difficulty levels based on data availability (distribution in Table 2). Each query includes a human-annotated reference plan with a rationale explaining the itinerary’s reasoning. By providing a structured evaluation framework with well-defined constraints and human-curated plans, TripCraft serves as a rigorous benchmark for measuring the reasoning and planning capabilities of language agents.

Days	Easy	Medium	Hard	Total
3-day	91	124	129	344
5-day	52	156	116	324
7-day	75	155	102	332

Table 2: TripCraft dataset distribution

3.2 Constraint and Persona Details

TripCraft integrates numerous constraints and persona components to enhance the evaluation of language agents in constraint-aware itinerary generation, as shown in Table 3. These improvements ensure that generated travel plans are realistic, well-structured, and aligned with user preferences.

Commonsense Constraints. In addition to the commonsense constraints in TravelPlanner, we introduce refinements to improve itinerary realism. First, the same event should not be repeated multiple times across a trip, ensuring diversity in experiences. Second, meal timings must have a minimum gap of four hours between breakfast, lunch, and dinner to maintain a natural schedule. Third, the point-of-interest (PoI) list must follow strict validity rules: each day’s itinerary must begin and end at the designated accommodation, except on the final day when the traveler departs. The list should be an ordered sequence of accommodations, attractions, and restaurants, ensuring adequate time gaps between flight arrivals and accommodation check-ins, as well as between accommodation check-outs and departures.

Hard Constraints. We introduce two new hard constraints to enhance itinerary structuring. First, each attraction belongs to one or more of 15 diverse categories (see Table 4), ensuring a well-distributed

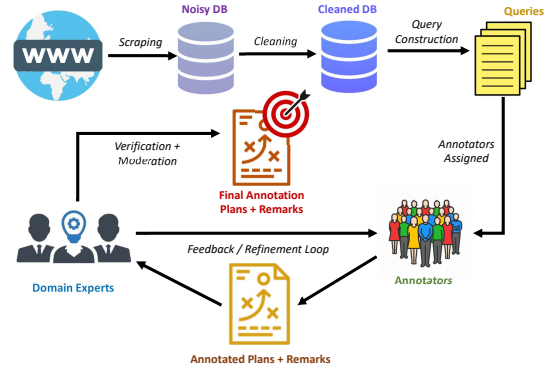


Figure 2: TripCraft dataset construction pipeline.

selection of activities. Second, events are categorized into four distinct types — Sports, Arts & Theatre, Music, and Film — allowing for a more structured and personalized planning process.

Persona Information. Each query in TripCraft is accompanied by a persona profile that influences travel planning, consisting of traveler type, purpose of travel, spending preference, and location preferences. Traveler type distinguishes between laid-back travelers, who prefer relaxation and scenic spots, and adventure seekers, who prioritize extreme activities like paragliding and bungee jumping. Purpose of travel refines traveler intent into categories such as relaxation, adventure, cultural exploration, and nature-focused experiences. Spending preference differentiates between luxury and budget-conscious travelers, shaping accommodation and activity choices. Location preferences specify favored destinations, such as beaches, mountains, cities, or wildlife/forest regions.

These persona aspects were carefully selected in consultation with domain experts as the most influential factors in determining a fulfilling travel experience. Unlike the work of Singh et al. (2024), which includes demographic attributes such as age, gender, and education, TripCraft deliberately omits such details to minimize potential biases in language model behavior.

3.3 Dataset Construction Pipeline

The benchmark construction (Figure 2) consists of three key steps as follows.

- **Data Collection and Cleaning:** The databases are sourced via web scraping and open-source tools like OSM² (details in Appendix A and Table 5). Given the extensive

²<https://www.openstreetmap.org/>

Commonsense Constraints	
Sufficient Meal Gaps	Defines a minimum gap of four hours between the meals.
Valid PoI list	Defines validity rules for the Point of Interest list.
Diverse Events	Event choices should not be repeated throughout the trip.
Hard Constraints	
Event Types	Event Types include four distinct categories—Sports, Arts & Theatre, Music, and Film.
Attraction Types	Each attraction belongs to one or more of 15 predefined categories, ensuring a well-distributed selection of activities.
Persona Components	
Traveler Type	Defines how a traveler approaches their journey—whether they seek relaxation in cozy spots or adrenaline-pumping adventures.
Purpose of Travel	Captures trip motivation. Examples: to unwind, explore cultures etc.
Spending Preference	Reflects the traveler’s budget and style, from luxurious indulgence to cost-conscious experiences.
Location Preference	Highlights preferred environments, such as beaches, mountains, cities, or wildlife-rich forests.

Table 3: Addition of constraints and persona details based on availability and limitations of scraped data has been given in the above table. The full list of constraints used in TripCraft has been given in Table 9 of Appendix.

size of the flights database, we adopt Xie et al. (2024)’s, adjusting dates to align with event timelines, ensuring queries incorporate relevant events while maintaining computational feasibility. Unlike prior datasets with outdated information, ours is more recent; however, missing or incomplete entries were either carefully removed or filled with default values to ensure consistency. Since GTFS public transit³ data covers only 140 cities, all databases are filtered accordingly.

- **Persona and Query Construction:** Following Xie et al. (2024), queries are generated by randomly selecting key elements—departure city, destination, and date range etc. Trip duration determines city coverage: 3-day plans focus on one city, while 5-day and 7-day plans span one state with visits to 2 and 3 cities, re-

³<https://gtfs.org/>

Category	Duration (hrs)
Boat Tours & Water Sports	3.5
Casinos & Gambling	2.5
Museums	3.0
Nature & Parks	4.5
Nightlife	2.5
Sights & Landmarks	3.0

Table 4: Attraction visiting duration (hrs) for a subset of categories. The complete list of categories has been given in Table 9 of Appendix A.

Database	Data Entries (#)
City Set	140
Flights	3,446,829
Restaurants	3,892
Attractions	5,043
Accommodations	2,800
Events	21,980
Nearest Transit Stop	8,723
Distance Matrix	19,460

Table 5: Data entries in the database have been scraped to include the most recent data.

spectively, requiring agents to reason about multi-city itineraries and inter-city connectivity. To enhance complexity, hard constraints and persona profiles are incorporated. These structured inputs are then composed using GPT-4o⁴ in a few-shot setting to generate high-quality queries.

- **Annotation and Refinement:** A team of 25 graduate students annotated plans for the queries⁵, providing justifications to enhance explainability. The process involved iterative refinements, integrating expert feedback to ensure nuanced interpretations of persona constraints. Domain Experts conducted a final manual review of all query-plan pairs, combining evaluation scripts with manual checks for feasibility and optimality. Annotation in

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

⁵Interns at our NLP lab.

Parameter	3-day	5-day	7-day
Restaurant Parameters			
Breakfast			
Mean Duration (hrs)	0.79	0.83	0.77
Std. Duration (hrs)	0.29	0.49	0.26
Lunch			
Mean Duration (hrs)	0.95	0.98	0.89
Std. Duration (hrs)	0.36	0.34	0.29
Dinner			
Mean Duration (hrs)	1.24	1.21	1.06
Std. Duration (hrs)	0.80	0.78	0.41
Attraction Parameters			
$\lambda_{laidback}$	1.24	1.27	1.14
$\lambda_{adventurous}$	1.91	1.64	1.78
σ_d (hrs)	1.09	0.98	0.96
n^{max}	5	4	4
n^{min}	0	0	0
k (hrs)	0.18	0.68	0.56

Table 6: Day-wise parameter details calculated from our annotation distribution. For a complete list of parameters refer to Table 10 in Appendix A.

TripCraft is notably more demanding (~30 minutes per instance) than prior datasets due to the added temporal and spatial complexities in the PoI list, requiring heightened scrutiny.

3.4 Evaluation Metrics

Plan feasibility is assessed using the hard and commonsense constraints from Subsection 3.2. Furthermore, we introduce continuous evaluation metrics that capture temporal, spatial, sequential, and persona-specific nuances of a travel plan, addressing the limitations of discrete binary checks (the advantages of these metrics are detailed in Subsection 4.2). The parameters of these metrics have been determined from our annotation distribution (Table 6). More generally, given annotations for any dataset, metric parameters can always be inferred from the distribution of corresponding variables in the annotation.

Temporal Meal Score: Assesses how well meal timings align with ideal breakfast, lunch, and dinner time intervals, ensuring a traveler enjoys meals without rushing. For each meal type (breakfast, lunch, or dinner), the temporal meal score for meal interval a , T_{meal}^a , is modeled using a bivariate normal distribution, motivated by prior work leveraging Gaussian models for time series data (Ansari et al., 2024; Eirola and Lendasse, 2013). This choice enables smooth penalization of deviations, ensuring a continuous representation of temporal

dependencies. Formally, T_{meal}^a is computed as:

$$T_{meal}^a = \frac{1}{(2\pi)\sqrt{\Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (1)$$

where $\mathbf{x} = [t_m, d_m]$ represents the midpoint time t_m and duration d_m of the meal, and Σ is the determinant of the covariance matrix with the correlation coefficient β . T_{meal}^a is normalized to obtain a score between 0 and 1 for each meal. Lastly, these values are averaged over all the meals in a plan:

$$\bar{T}_{meal} = \frac{1}{M} \sum_{a=1}^M T_{meal}^a \quad (2)$$

where M is the total number of meals in the plan. **Temporal Attraction Score:** Evaluates whether the visit duration of an attraction aligns with an expected duration for that type of attraction while also considering the total number of attractions visited in a day. It is modeled as the probability density function of a joint mixed random distribution combining a Gaussian for duration (X) and a Poisson for the number of attractions (N). The Poisson distribution is widely used for modeling discrete variables representing independent event counts within a fixed time or space interval (Hu, 2008; Inouye et al., 2017). This makes it suitable for scenarios such as modeling the number of attractions visited by tourists:

$$f_{X,N}(d_i, n) = f_{X|N}(d_i | n) \cdot P(N = n) \quad (3)$$

Averaging over all attractions, n , in a plan:

$$\bar{T}_{attrac} = \frac{1}{n} \sum_{i=1}^n \exp\left(-\frac{(d_i - \mu_d^i)^2}{2\sigma_d^2}\right) \cdot \frac{\lambda^n e^{-\lambda}}{n!} \quad (4)$$

Here, d_i represents the actual duration spent at attraction i , while μ_d^i denotes the expected visit duration for i , which is dynamically adjusted based on the number of attractions visited and attraction type. σ_d captures variability in visit durations. To incorporate the influence of the total number of attractions, we use a Poisson distribution P whose λ parameter represents the expected number of attractions for a given persona (e.g., adventurous or laidback). The expected visit duration μ_d^i is adjusted dynamically:

$$\mu_d^i = \begin{cases} \mu_d^{\text{type}(i)} - k(n - n^{\min}), & \text{if adventurous} \\ \mu_d^{\text{type}(i)} + k(n^{\max} - n), & \text{if laidback} \end{cases} \quad (5)$$

where $\mu_d^{\text{type}(i)}$ represents the expected duration for the category of attraction i , and k denotes the rate

at which the average visit duration changes per unit increase in the number of attractions. Intuitively, for a particular type of attraction, a laid-back traveler generally spends more time compared to an adventure seeker. Moreover, irrespective of the traveler’s persona, the average duration spent at each attraction decreases as the total number of attractions visited in a day increases.

Spatial Score: Measures the proximity of public transit stops to generated POIs; closer stops indicate greater travel convenience. It is computed based on the distance d of a point of interest from the nearest transit station:

$$S_s(d) = \begin{cases} 1 - 0.5 \left(\frac{d}{d_0} \right), & \text{if } d \leq d_0 \\ 0.5 \exp(-\lambda(d - d_0)), & \text{if } d > d_0 \end{cases} \quad (6)$$

where d_0 ($= 5$ km) is a threshold distance and λ ($= 0.0002$) is the decay rate for larger distances. The final spatial score for a plan is the average of all individual POI scores:

$$\bar{S}_{spatial} = \frac{1}{N} \sum_{i=1}^N S_s(d_i), \quad N \text{ is the no. of POIs visited.} \quad (7)$$

Persona Score: Quantifies the alignment between a traveler’s persona and visited POIs. Prior work on POI conflation (Sun et al., 2023) explored type- and name-based methods; we prioritize POI names as they offer richer semantics, often embedding key descriptors like *relaxing* or *luxury*. The persona score is computed as the average cosine similarity between BERT (Devlin et al., 2019) embeddings of persona components and POI names:

$$\bar{S}_{persona} = \frac{1}{M \cdot N} \sum_{j=1}^M \sum_{i=1}^N \frac{\mathbf{p}_j \cdot \mathbf{q}_i}{\|\mathbf{p}_j\| \|\mathbf{q}_i\|} \quad (8)$$

where \mathbf{p}_j represents the BERT embedding of the j -th persona component, \mathbf{q}_i is the BERT embedding of the i -th POI name, M is the total number of persona components, and N is the total number of POIs in the travel plan.

Ordering Score: Measures the sequential alignment of the generated POI list with the annotated POI list for a particular day:

$$S_{ord} = 1 - \frac{ED(\mathcal{G}, \mathcal{A})}{\max(|\mathcal{G}|, |\mathcal{A}|)} \quad (9)$$

where \mathcal{G} and \mathcal{A} represent the generated and annotated sequences of POIs, respectively, and $ED(\mathcal{G}, \mathcal{A})$ denotes the Edit Distance (Levenshtein, 1965) between them. This is averaged over all the days of the journey to get the ordering score for a

plan, which is denoted by \bar{S}_{ord} . It evaluates how closely the sequence of POIs in the generated plan matches the annotated plan, penalizing missing or misordered POIs. For instance, if the annotated plan contains the sequence breakfast \rightarrow attraction \rightarrow lunch, but the generated plan skips the attraction or clusters all attractions post-lunch, the edit distance increases, lowering the ordering score.

4 Experiments and Results

4.1 Experimental Settings

We adopt the direct sole planning strategy (Xie et al., 2024; Singh et al., 2024), modifying the prompt to include event details, POI lists, and a refined one-shot example tailored to our constraints. These enhancements improve alignment with the domain-specific requirements. Additionally, we introduce a novel setting, direct sole planning (with parameter information), which extends the prompt with natural language descriptions of the parameterized distributions modeling the metrics in Subsection 3.4. Appendix B and Figure 3 illustrate the distinction between these settings.

Incorporating parameter information encourages LLMs to adhere to precise timings and contextual constraints beyond subjective commonsense reasoning. For example, while breakfast is typically in the morning, its exact timing is ambiguous. Since our framework evaluates plans using continuous probability distributions, explicit constraints ensure fairness and optimization for LLM-based planning.

Existing travel-specific models like TOURISM-NLG (Bhatt et al., 2023) are much smaller than LLMs. Hence, we evaluate GPT-4o, Qwen2.5-7B-Instruct, and Phi-4-mini-Instruct across 3-day, 5-day, and 7-day travel plans, comparing both settings using proposed and existing metrics⁶. For fairness, we report our proposed metrics by averaging results over the set of travel plans that got delivered in both settings (i.e., the intersection of their delivered plans). Results are summarized in Tables 7 and 8 with key insights in next subsection.

4.2 Discussions

This section discusses the observations from the experiments, the advantages of our proposed evaluation metrics, and highlights key challenges posed by our dataset.

⁶Since we are not training/ fine-tuning LLMs, we directly report results without making train-val splits.

Model	Setting	Category	\bar{T}_{meal}	\bar{T}_{attrac}	$\bar{S}_{spatial}$	$\bar{S}_{persona}$	\bar{S}_{ord}
GPT-4o	w/o Parameter Info	3-day	0.56	0.0169	0.80	0.46	0.70
		5-day	0.69	0.0057	0.86	0.49	0.91
		7-day	0.61	0.0122	0.85	0.50	0.96
	w/ Parameter Info	3-day	0.70	0.0171	0.83	0.48	0.74
		5-day	0.75	0.0078	0.87	0.50	0.93
		7-day	0.80	0.0057	0.86	0.51	0.97
Qwen2.5-7B-Instruct	w/o Parameter Info	3-day	0.53	0.0233	0.834	0.512	0.769
		5-day	0.55	0.0083	0.667	0.505	0.918
		7-day	0.52	0.0098	0.658	0.516	0.956
	w/ Parameter Info	3-day	0.58	0.0150	0.869	0.514	0.792
		5-day	0.59	0.0085	0.742	0.509	0.924
		7-day	0.57	0.0100	0.734	0.524	0.962
Phi-4-mini-Instruct	w/o Parameter Info	3-day	0.228	0.018	0.62	0.519	0.56
		5-day	0.508	0.006	0.82	0.525	0.91
		7-day	0.476	0.017	0.81	0.527	0.95
	w/ Parameter Info	3-day	0.244	0.022	0.60	0.529	0.66
		5-day	0.526	0.013	0.83	0.526	0.92
		7-day	0.505	0.014	0.83	0.531	0.96

Table 7: Results of our proposed evaluation metrics for the two settings (w/ or w/o Parameter Info) with 3 LLMs. For each category, the **bolded** values represent the better ones between the two settings.

Model	Setting	Category	Delivery	CPR		HCPR		Final
			Rate	Micro	Macro	Micro	Macro	Pass Rate
GPT-4o	w/o Parameter Info	3-day	92.60	77.61	6.08	27.63	26.08	3.47
		5-day	98.69	66.99	2.17	4.21	3.04	1.74
		7-day	96.03	70.35	<u>0.00</u>	0.62	<u>0.00</u>	<u>0.00</u>
	w/ Parameter Info	3-day	96.08	80.08	5.21	26.10	25.21	1.74
		5-day	91.30	61.69	0.87	8.60	5.21	0.43
		7-day	92.51	66.43	<u>0.00</u>	0.83	<u>0.00</u>	<u>0.00</u>
Qwen2.5-7B-Instruct	w/o Parameter Info	3-day	100	72.04	0.0	<u>3.26</u>	3.04	<u>0.0</u>
		5-day	100	46.30	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
		7-day	<u>99.13</u>	40.04	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
	w/ Parameter Info	3-day	99.56	70.39	0.43	<u>3.26</u>	2.60	<u>0.0</u>
		5-day	99.56	52.30	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
		7-day	<u>99.13</u>	39.29	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
Phi-4-mini-Instruct	w/o Parameter Info	3-day	95.21	49.73	<u>0.0</u>	0.38	0.86	<u>0.0</u>
		5-day	98.69	43.04	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
		7-day	<u>97.79</u>	37.84	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
	w/ Parameter Info	3-day	92.60	47.69	<u>0.0</u>	0.0	0.0	<u>0.0</u>
		5-day	99.56	43.86	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>
		7-day	<u>97.79</u>	37.22	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>	<u>0.0</u>

Table 8: Results of existing evaluation metrics for the two settings (w/ or w/o Parameter Info) with 3 LLMs. CPR and HCPR stand for Commonsense Pass Rate and Hard Constraint Pass Rate respectively. For each category, the **bolded** values represent the better ones between the two settings. Moreover, equal values have been underlined.

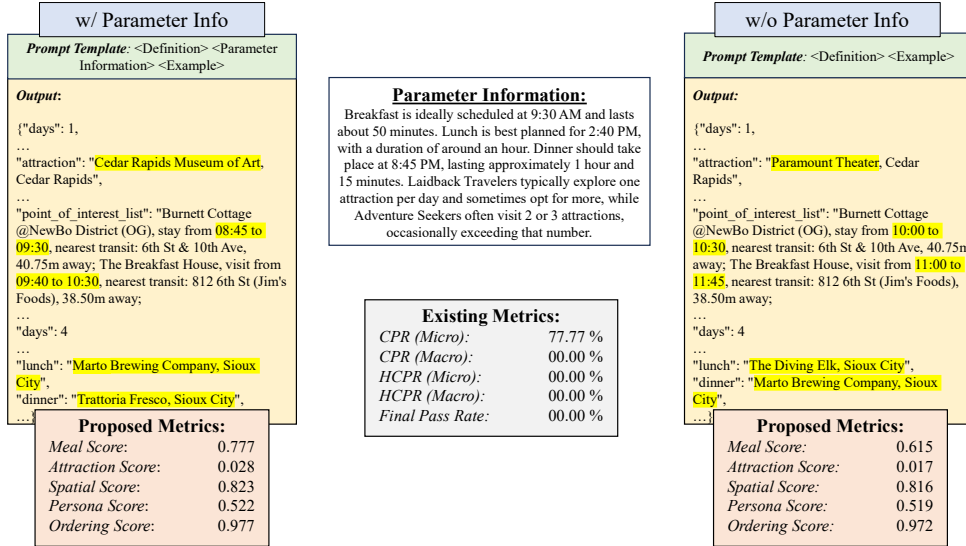


Figure 3: Existing constraint-based metrics rate these plans equally, but, our metrics provide a continuous score, capturing temporal, spatial, sequential, and personal differences.

Observation 1. Trade-off between objective metrics and constraint adherence: Table 7 shows that incorporating parameter information significantly improves objective metrics, particularly in meal scheduling and spatial efficiency. However, this improvement comes at the cost of increased constraint violations, as evidenced by lower commonsense pass rates in Table 8. This trade-off suggests that while parameter-informed guidance aids LLMs in structuring itineraries, it may also introduce rigid assumptions that conflict with real-world constraints, requiring future methods to balance optimization and feasibility effectively.

Observation 2. Potential of objective criteria: Traditional percentage-based evaluations fail to distinguish between travel plans of equal feasibility under given constraints. In Table 8, the 7-day travel plans receive a zero macro pass rate for commonsense and hard constraints, yet our proposed metrics differentiate plans based on aspects beyond binary pass/fail outcomes. As illustrated in Figure 3, our proposed continuous scoring metric captures the difference, such as partial alignment with temporal constraints and near-optimal spatial organization, providing a more interpretable assessment of LLM-generated plans.

Observation 3. Limitations in LLM-Generated Itineraries: Despite structured prompts, the LLM agent exhibits temporal inconsistencies, where PoI visits extend beyond departure times, meal schedules fail to align with natural dining hours, and activity timestamps are misordered, disrupting

itinerary flow (Refer to Appendix C). Additionally, agents struggle to adapt activity density to different personas, often under-scheduling for adventure-seeking travelers while generating excessive transit between locations for relaxed travelers. Another limitation is spatial reasoning, where transit assignments are often impractical, with accommodations and attractions placed far from available transit stops, leading to infeasible travel plans. These issues suggest that while LLMs can incorporate structured constraints, they lack robust reasoning over real-world travel logistics, user preferences, and sequential dependencies, making TripCraft a crucial benchmark for future advancements.

5 Conclusion

In this work, we introduce TripCraft, a high-fidelity travel planning dataset that surpasses prior benchmarks in realism and consistency. It ensures spatiotemporal coherence while integrating fine-grained attraction categories, persona components, event typologies, and transit schedules for personalized itinerary planning. Beyond dataset improvements, we enhance itinerary generation with higher spatiotemporal resolution and structured PoI sequencing. To evaluate LLM-generated plans, we introduce mathematical scoring paradigms that complement constraint-based validation. Through TripCraft and our novel evaluation framework, we set a new standard for AI-driven travel planning.

Limitations

While TripCraft significantly enhances the realism and coherence of travel planning datasets, certain limitations remain. One key constraint is the exclusion of events from the core Point of Interest (PoI) list due to the lack of structured temporal information. If such data becomes available, it can be incorporated into our dataset, enabling LLMs to explicitly factor in event timings, ultimately leading to more mature and contextually aware itineraries.

Our dataset is currently designed for U.S. cities, but the construction pipeline can be extended to other geographical regions if the necessary data is available. Expanding to a global scale would require accounting for region-specific differences in travel preferences, transportation infrastructure, cultural factors, as well as diverse annotator demographics, which remain open challenges for future research.

Additionally, our primary objective is not to propose a novel travel planning methodology but to establish a robust dataset, enhanced itinerary generation, and fine-grained evaluation metrics. While TripCraft provides a strong benchmark for assessing LLM-driven travel planning, future research may explore diverse methodologies on this dataset, further advancing personalized AI-driven itinerary synthesis.

Ethics Statement

Our study utilizes publicly available web data, which we have carefully scraped to construct our databases while ensuring compliance with relevant terms of use and ethical considerations. To safeguard privacy, we have fully anonymized sensitive personal details. However, with annotators' consent, aggregate demographic statistics are provided in Appendix E. We are committed to transparency and reproducibility in research; therefore, we release both the datasets and code to facilitate further exploration by the scientific community.

Acknowledgments

This work was partially supported by the Microsoft Academic Partnership Grant (MAPG) 2024. We sincerely thank the annotators for their diligent and meticulous efforts in curating and labeling the data, which were essential for the creation of the dataset. We also thank the anonymous reviewers for their valuable feedback and constructive sugges-

tions, which helped enhance the quality and clarity of the paper.

References

- Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. 2024. [Chronos: Learning the language of time series](#). *Transactions on Machine Learning Research*. Expert Certification.
- Sahil Manoj Bhatt, Sahaj Agarwal, Omkar Gurjar, Manish Gupta, and Manish Shrivastava. 2023. [Tourismnlg: a multi-lingual generative benchmark for the tourism domain](#). In *European Conference on Information Retrieval*, pages 150–166. Springer.
- 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 preprint arXiv:2406.13094*.
- Andrey Borro, Patricia J Riddle, Michael W Barley, and Michael J Witbrock. 2025. [Large language models as common-sense heuristics](#). *arXiv preprint arXiv:2501.18816*.
- Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. 2024. [Travelagent: An ai assistant for personalized travel planning](#). *Preprint, arXiv:2409.08069*.
- Rémi Coulom. 2006. [Efficient selectivity and backup operators in monte-carlo tree search](#). In *Proceedings of the 5th International Conference on Computers and Games, CG'06*, page 72–83, Berlin, Heidelberg. Springer-Verlag.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emil Eirola and Amaury Lendasse. 2013. [Gaussian mixture models for time series modelling, forecasting, and interpolation](#). In *Advances in Intelligent Data Analysis XII*, pages 162–173, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Atharva Gundawar, Mudit Verma, Lin Guan, Karthik Valmееkam, 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.
- Hao Hu. 2008. Poisson distribution and application. [A Course in Department of Physics and Astronomy; University of Tennessee at Knoxville: Knoxville, TN, USA](#).
- Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. 2024. Making large language models better planners with reasoning-decision alignment. In [European Conference on Computer Vision](#), pages 73–90. Springer.
- David I Inouye, Eunho Yang, Genevera I Allen, and Pradeep Ravikumar. 2017. A review of multivariate distributions for count data derived from the poisson distribution. [Wiley Interdisciplinary Reviews: Computational Statistics](#), 9(3):e1398.
- 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.
- Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. Evolving deeper llm thinking. [arXiv preprint arXiv:2501.09891](#).
- Vladimir I. Levenshtein. 1965. [Binary codes capable of correcting deletions, insertions, and reversals](#). [Soviet physics. Doklady](#), 10:707–710.
- Vishal Pallagani, Bharath Muppasani, Keerthiram Murgesan, Francesca Rossi, Biplav Srivastava, Lior Horesh, Francesco Fabiano, and Andrea Loreggia. 2023. Understanding the capabilities of large language models for automated planning. [arXiv preprint arXiv:2305.16151](#).
- Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2024. Adapt: As-needed decomposition and planning with language models. In [Findings of the Association for Computational Linguistics: NAACL 2024](#), pages 4226–4252.
- 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](#).
- 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 \(ICCV\)](#).
- Kai Sun, Yingjie Hu, Yue Ma, Ryan Zhenqi Zhou, and Yunqiang Zhu. 2023. Conflating point of interest (poi) data: A systematic review of matching methods. [Computers, Environment and Urban Systems](#), 103:101977.
- 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, et al. 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, et al. 2025. The rise and potential of large language model based agents: A survey. [Science China Information Sciences](#), 68(2):121101.
- 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.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023.

React: Synergizing reasoning and acting in language models. Preprint, arXiv:2210.03629.

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.

Overview of Appendices

This supplementary material presents additional details on the following aspects:

- **Appendix A:** Data Sourcing Details
- **Appendix B:** Prompt and Annotation Details
- **Appendix C:** Case Studies
- **Appendix D:** Impact of Distribution Shift
- **Appendix E:** Annotator Details

A Data Sourcing Details

Our dataset is constructed using current data sources to ensure spatio-temporal consistency and personalization. Below, we detail the sourcing methodology and heuristics for each component.

A.1 Restaurants

We extracted restaurant details using **TripAdvisor’s Apify scraper**⁷, which provided all necessary attributes except precise pricing. TripAdvisor denotes cost using dollar symbols (\$–\$\$\$) instead of exact values. To estimate absolute prices, we leveraged city-specific restaurant price indices from **Numbeo**⁸, scaling them according to the number of dollar symbols in each price rating.

Category	Duration (hrs)
Boat Tours & Water Sports	3.5
Casinos & Gambling	2.5
Classes & Workshops	1.5
Concerts & Shows	2.5
Food & Drink	2.5
Fun & Games	1.5
Museums	3.0
Nature & Parks	4.5
Nightlife	2.5
Outdoor Activities	4.0
Shopping	1.5
Sights & Landmarks	3.0
Spas & Wellness	2.0
Water & Amusement Parks	5.0
Zoos & Aquariums	2.5

Table 9: Attraction visiting duration (hrs) for each category. Note that an attraction can belong to one or more than one categories.

A.2 Attractions

Attraction details, including subcategories, were sourced from **TripAdvisor’s Apify scraper**⁷.

⁷<https://console.apify.com/actors/dbEyMBriog95Fv8CW/input>

⁸<https://www.numbeo.com/cost-of-living/>

Parameter	3-day	5-day	7-day
Restaurant Parameters			
Breakfast			
Mean Time	9.55	9.57	9.67
Mean Duration (hrs)	0.79	0.83	0.77
Std. Time	1.19	0.75	0.80
Std. Duration (hrs)	0.29	0.49	0.26
Beta	0.00	0.19	-0.18
Lunch			
Mean Time	14.62	14.61	14.69
Mean Duration (hrs)	0.95	0.98	0.89
Std. Time	0.99	0.69	0.81
Std. Duration (hrs)	0.36	0.34	0.29
Beta	-0.03	0.01	0.15
Dinner			
Mean Time	20.73	20.86	20.74
Mean Duration (hrs)	1.24	1.21	1.06
Std. Time	1.31	1.23	1.24
Std. Duration (hrs)	0.80	0.78	0.41
Beta	-0.18	-0.09	-0.25
Attraction Parameters			
λ_{aidback}	1.24	1.27	1.14
$\lambda_{\text{adventurous}}$	1.91	1.64	1.78
σ_d (hrs)	1.09	0.98	0.96
n^{max}	5	4	4
n^{min}	0	0	0
k (hrs)	0.18	0.68	0.56

Table 10: A comprehensive list of parameter details for 3-day, 5-day, and 7-day scenarios as calculated from the annotation distribution statistics.

Since a majority of attractions lacked predefined visit durations, we consulted domain experts to establish category-wise average durations for each attraction type. Finally, each attraction’s duration was assigned as the mean of the categories it belonged to, ensuring a realistic time allocation (Table 9).

A.3 Flights

We adopted the flight database from (Xie et al., 2024) but adjusted all dates to November 2024 to maximize temporal alignment with event data. This adjustment ensures that LLM-generated itineraries incorporate relevant event-based recommendations.

A.4 Distance Matrices

All pairwise distances were computed using **OpenStreetMap’s OSRM API**⁹, ensuring accurate and real-time routing information.

⁹<http://project-osrm.org/>

A.5 Accommodations

We scraped accommodation listings from Airbnb using **Apify’s Airbnb scraper**¹⁰. Since minimum stay requirements were not available in the extracted data, we excluded this attribute from our dataset.

A.6 Events

Event data was collected using **Ticketmaster’s Apify scraper**¹¹, covering a diverse range of concerts, sports, theater, and other entertainment events.

A.7 Public Transit

We sourced transit schedules from the **General Transit Feed Specification (GTFS)**¹² for 140 cities. For each Point of Interest (PoI)—including accommodations, restaurants, and attractions—we determined the nearest public transit stop using geodesic distance (computed via **Geopy**¹³). This enables LLMs to incorporate realistic public transit connectivity when generating travel itineraries.

¹⁰<https://apify.com/dtrungtin/airbnb-scraper>

¹¹<https://console.apify.com/actors/H17bNMx0vqaqvdfZQ>

¹²<https://gtfs.org/>

¹³<https://pypi.org/project/geopy/>

Constraint	Description
<i>Environment Constraint</i>	
Unavailable Transportation	There is no available flight or driving information between the two cities.
Unavailable Attractions	There is no available attraction information in the queried city.
<i>Commonsense Constraint</i>	
Within Sandbox	All information in the plan must be within the closed sandbox; otherwise, it will be considered a hallucination.
Complete Information	No key information should be left out of the plan, such as the lack of accommodation during travel.
Sufficient Meal Gaps (+)	Meal timings must have a minimum gap of four hours between breakfast, lunch, and dinner to maintain a natural schedule.
Valid PoI list (+)	The point-of-interest (PoI) list must follow strict validity rules: each day’s itinerary must begin and end at the designated accommodation, except on the final day when the traveler departs. The list should be limited to accommodations, attractions, and restaurants, ensuring adequate time gaps between flight arrivals and accommodation check-ins, as well as between accommodation check-outs and departures.
Diverse Events (+)	Event choices should not be repeated throughout the trip.
Within Current City	All scheduled activities for the day must be located within that day’s city(ies).
Reasonable City Route	Changes in cities during the trip must be reasonable.
Diverse Restaurants	Restaurant choices should not be repeated throughout the trip.
Diverse Attractions	Attraction choices should not be repeated throughout the trip.
Non-conf. Transportation	Transportation choices within the trip must be reasonable. For example, having both “self-driving” and “flight” would be considered a conflict.
<i>Hard Constraint</i>	
Budget	The total budget of the trip.
Room Rule	Room rules include “No parties”, “No smoking”, “No children under 10”, “No pets”, and “No visitors”.
Room Type	Room types include “Entire Room”, “Private Room”, “Shared Room”, and “No Shared Room”.
Cuisine	Cuisines include “Chinese”, “American”, “Italian”, “Mexican”, “Indian”, “Mediterranean”, and “French”.
Transportation	Transportation options include “No flight” and “No self-driving”.
Event Types (+)	Event Types include four distinct categories—Sports, Arts & Theatre, Music, and Film.
Attraction Types (+)	Each attraction belongs to one or more of 15 predefined categories, ensuring a well-distributed selection of activities.
<i>Persona Components</i>	
Traveler Type (+)	Defines how a traveler approaches their journey—whether they seek relaxation in cozy spots or adrenaline-pumping adventures.
Purpose of Travel (+)	Captures the main motivation behind the trip, whether it’s to unwind, seek thrills, explore cultures, or connect with nature.
Spending Preference (+)	Reflects the traveler’s budget and style, from luxurious indulgence to cost-conscious experiences.
Location Preference (+)	Highlights preferred environments, such as beaches, mountains, cities, or wildlife-rich forests.

Table 11: *Comprehensive Constraint and Persona Description.* (+) denotes the ones we have added.

B Prompt and Annotation Details

B.1 Prompt with Parameter Info

You are a proficient planner. Based on the provided information, query and persona, please give a detailed travel plan, including specifics such as flight numbers (e.g., F0123456), restaurant names, and accommodation names. Note that all the information in your plans should be derived from the provided data. You must adhere to the format given in the example. Additionally, all details should align with common sense. The symbol '-' indicates that information is unnecessary. For example, in the provided sample, you do not need to plan after returning to the departure city. When you travel to two cities in one day, you should note it in the "Current City" section as in the example (i.e., from A to B). Include events happening on that day, if any. Provide a Point of Interest List, which is an ordered list of places visited throughout the day. This list should include accommodations, attractions, or restaurants and their starting and ending timestamps. Each day must start and end with the accommodation where the traveler is staying. Breakfast is ideally scheduled at 9:30 AM and lasts about 50 minutes. Lunch is best planned for 2:40 PM, with a duration of around an hour. Dinner should take place at 8:45 PM, lasting approximately 1 hour and 15 minutes. Laidback Travelers typically explore one attraction per day and sometimes opt for more, while Adventure Seekers often visit 2 or 3 attractions, occasionally exceeding that number.

***** Example *****

Query: Could you create a travel plan for 7 people from Ithaca to Charlotte spanning 3 days, from March 8th to March 14th, 2022, with a budget of \$30,200?

Traveler Persona:

Traveler Type: Laidback Traveler;

Purpose of Travel: Relaxation;

Spending Preference: Economical Traveler;

Location Preference: Beaches

Travel Plan:

Day 1:

Current City: from Ithaca to Charlotte

Transportation: Flight Number: F3633413, from Ithaca to Charlotte, Departure Time: 05:15, Arrival Time: 07:28

Breakfast: Nagaland's Kitchen, Charlotte

Attraction: The Charlotte Museum of History, Charlotte

Lunch: Cafe Maple Street, Charlotte

Dinner: Bombay Vada Pav, Charlotte

Accommodation: Affordable Spacious Refurbished Room in Bushwick!, Charlotte

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 08:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Nagaland's Kitchen, visit from 09:00 to 09:45, nearest transit: Uptown Station, 200m away; The Charlotte Museum of History, visit from 10:30 to 13:30, nearest transit: Museum Station, 300m away; Cafe Maple Street, visit from 14:00 to 15:00, nearest transit: Maple Avenue Stop, 100m away; Bombay Vada Pav, visit from 19:00 to 20:00, nearest transit: Bombay Stop, 150m away; Affordable Spacious Refurbished Room in Bushwick!, stay from 21:00 to 07:00, nearest transit: Bushwick Stop, 100m away.

Day 2:

Current City: Charlotte

Transportation: -

Breakfast: Olive Tree Cafe, Charlotte

Attraction: The Mint Museum, Charlotte; Romare Bearden Park, Charlotte

Lunch: Birbal Ji Dhaba, Charlotte

Dinner: Pind Balluchi, Charlotte

Accommodation: Affordable Spacious Refurbished Room in Bushwick!, Charlotte

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 07:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Olive Tree Cafe, visit from 09:00 to 09:45, nearest transit: Cafe Station, 250m away; The Mint Museum, visit from 10:30 to 13:00, nearest transit: Mint Stop, 200m away; Birbal Ji Dhaba, visit from 14:00 to 15:30, nearest transit: Dhaba Stop, 120m away; Romare Bearden Park, visit from 16:00 to 18:00, nearest transit: Park Stop, 150m away; Pind Balluchi, visit from 19:30 to 21:00, nearest transit: Pind Stop, 150m away; Affordable Spacious Refurbished Room in Bushwick!, stay from 21:30 to 07:00, nearest transit: Bushwick Stop, 100m away.

Day 3:

Current City: from Charlotte to Ithaca

Transportation: Flight Number: F3786167, from Charlotte to Ithaca, Departure Time: 21:42, Arrival Time: 23:26

Breakfast: Subway, Charlotte

Attraction: Books Monument, Charlotte

Lunch: Olive Tree Cafe, Charlotte

Dinner: Kylin Skybar, Charlotte

Accommodation: -

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 07:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Subway, visit from 09:00 to 10:00, nearest transit: Subway Station, 150m away; Books Monument, visit from 10:30 to 13:30, nearest transit: Central Library Stop, 200m away; Olive Tree Cafe, visit from 14:00 to 15:00, nearest transit: Cafe Station, 250m away; Kylin Skybar, visit from 19:00 to 20:00, nearest transit: Skybar Stop, 180m away.

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

Given information: {text}

Query: {query}

Traveler Persona:

{persona}

Output:

B.2 Prompt without Parameter Info

You are a proficient planner. Based on the provided information, query and persona, please give a detailed travel plan, including specifics such as flight numbers (e.g., F0123456), restaurant names, and accommodation names. Note that all the information in your plans should be derived from the provided data. You must adhere to the format given in the example. Additionally, all details should align with common sense. The symbol '-' indicates that information is unnecessary. For

example, in the provided sample, you do not need to plan after returning to the departure city. When you travel to two cities in one day, you should note it in the "Current City" section as in the example (i.e., from A to B). Include events happening on that day, if any. Provide a Point of Interest List, which is an ordered list of places visited throughout the day. This list should include only accommodations, attractions, or restaurants and their starting and ending timestamps. Each day must start and end with the accommodation where the traveler is staying.

***** Example *****

Query: Could you create a travel plan for 7 people from Ithaca to Charlotte spanning 3 days, from March 8th to March 14th, 2022, with a budget of \$30,200?

Traveler Persona:

Traveler Type: Laidback Traveler;

Purpose of Travel: Relaxation;

Spending Preference: Economical Traveler;

Location Preference: Beaches

Travel Plan:

Day 1:

Current City: from Ithaca to Charlotte

Transportation: Flight Number: F3633413, from Ithaca to Charlotte, Departure Time: 05:15, Arrival Time: 07:28

Breakfast: Nagaland's Kitchen, Charlotte

Attraction: The Charlotte Museum of History, Charlotte

Lunch: Cafe Maple Street, Charlotte

Dinner: Bombay Vada Pav, Charlotte

Accommodation: Affordable Spacious Refurbished Room in Bushwick!, Charlotte

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 08:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Nagaland's Kitchen, visit from 09:00 to 09:45, nearest transit: Uptown Station, 200m away; The Charlotte Museum of History, visit from 10:30 to 13:30, nearest transit: Museum Station, 300m away; Cafe Maple Street, visit from 14:00 to 15:00, nearest transit: Maple Avenue Stop, 100m away; Bombay Vada Pav, visit from 19:00 to 20:00, nearest transit: Bombay Stop, 150m away; Affordable Spacious Refurbished Room in Bushwick!, stay from 21:00 to 07:00, nearest transit: Bushwick Stop, 100m away.

Day 2:

Current City: Charlotte

Transportation: -

Breakfast: Olive Tree Cafe, Charlotte

Attraction: The Mint Museum, Charlotte; Romare Bearden Park, Charlotte

Lunch: Birbal Ji Dhaba, Charlotte

Dinner: Pind Balluchi, Charlotte

Accommodation: Affordable Spacious Refurbished Room in Bushwick!, Charlotte

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 07:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Olive Tree Cafe, visit from 09:00 to 09:45, nearest transit: Cafe Station, 250m away; The Mint Museum, visit from 10:30 to 13:00, nearest transit: Mint Stop, 200m away; Birbal Ji Dhaba, visit from 14:00 to 15:30, nearest transit: Dhaba Stop, 120m away; Romare Bearden Park, visit from 16:00 to 18:00, nearest transit: Park Stop,

150m away; Pind Balluchi, visit from 19:30 to 21:00, nearest transit: Pind Stop, 150m away; Affordable Spacious Refurbished Room in Bushwick!, stay from 21:30 to 07:00, nearest transit: Bushwick Stop, 100m away.

Day 3:

Current City: from Charlotte to Ithaca

Transportation: Flight Number: F3786167, from Charlotte to Ithaca, Departure Time: 21:42, Arrival Time: 23:26

Breakfast: Subway, Charlotte

Attraction: Books Monument, Charlotte

Lunch: Olive Tree Cafe, Charlotte

Dinner: Kylin Skybar, Charlotte

Accommodation: -

Event: -

Point of Interest List: Affordable Spacious Refurbished Room in Bushwick!, stay from 07:00 to 08:30, nearest transit: Bushwick Stop, 100m away; Subway, visit from 09:00 to 10:00, nearest transit: Subway Station, 150m away; Books Monument, visit from 10:30 to 13:30, nearest transit: Central Library Stop, 200m away; Olive Tree Cafe, visit from 14:00 to 15:00, nearest transit: Cafe Station, 250m away; Kylin Skybar, visit from 19:00 to 20:00, nearest transit: Skybar Stop, 180m away.

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

Given information: {text}

Query: {query}

Traveler Persona:

{persona}

Output:

B.3 Annotation and Remark (by Human Annotator)

Day 1:

Current City: from Charlotte to Houston

Transportation: Flight Number: F0180769, from Charlotte to Houston, Departure Time: 20:04, Arrival Time: 21:46

Breakfast: -

Attraction: -

Lunch: -

Dinner: -

Accommodation: Entire Apt in the Heart of the City - Galleria, Houston

Event: -

Point of Interest List:

Entire Apt in the Heart of the City - Galleria, stay from 22:30 to 08:00, nearest transit: Dr @ Burgoyne Rd, 98.06m away.

Day 2:

Current City: Houston

Transportation: -

Breakfast: Niko Niko's Greek

Attraction: Houston Zoo; Minute Maid Park; Bayou Bend Collection and Gardens

Lunch: Hugo's

Dinner: The Original Ninfa's

Accommodation: Entire Apt in the Heart of the City - Galleria, Houston

Event: -

Point of Interest List:

Entire Apt in the Heart of the City - Galleria, stay from 08:00 to 09:00, nearest transit: Dr @ Burgoyne Rd, 98.06m away;

Niko Niko's Greek, visit from 09:30 to 10:30, nearest transit: Greek Montrose Blvd @ Missouri St, 22.62m away;

Houston Zoo, visit from 11:00 to 13:00, nearest transit: Cambridge St @ Ben Taub Loop, 473.45m away;

Hugo's, visit from 13:30 to 14:30, nearest transit: Westheimer Rd @ Mandell St, 26.11m away;

Minute Maid Park, visit from 15:00 to 17:00, nearest transit: Park Congress St @ Jackson St, 170.01m away;

Bayou Bend Collection and Gardens, visit from 17:30 to 19:30, nearest transit: Memorial Dr @ Knox St, 87.75m away;

The Original Ninfa's, visit from 20:30 to 21:30, nearest transit: Canal St @ N Delano St, 153.30m away;

Entire Apt in the Heart of the City - Galleria, stay from 22:00 to 06:00, nearest transit: Dr @ Burgoyne Rd, 98.06m away.

Day 3:

Current City: from Houston to Charlotte

Transportation: Flight Number: F1462842, from Houston to Charlotte, Departure Time: 08:15, Arrival Time: 12:00

Breakfast: Phoenicia Specialty Foods

Attraction: -

Lunch: -

Dinner: -

Accommodation: Entire Apt in the Heart of the City - Galleria, Houston

Event: -

Point of Interest List:

Entire Apt in the Heart of the City - Galleria, stay from 06:00 to 07:00, nearest transit: Dr @ Burgoyne Rd, 98.06m away;

Phoenicia Specialty Foods, visit from 7:15 to 7:45, nearest transit: Lamar St @ Austin St, 9.87m away.

Remark

The food is within budget and aligns with the given preferences. The accommodation also meets the specified requirements. The theme is nature, as nature-related activities were preferred, while concerts were excluded from the itinerary.

B.4 Query Generation using GPT-4o

Given a JSON, please help me generate a natural language query. In the JSON, 'org' denotes the departure city. When 'days' exceeds 3, 'visiting_city_number' specifies the number of cities to be covered in the destination state. Please disregard the 'level' attribute. Here are three examples.

—EXAMPLE 1—

JSON:

```
{"org": "Washington", "dest": "Atlanta", "days": 3, "visiting_city_number": 1, "date": ["2024-11-18", "2024-11-19", "2024-11-20"], "people_number": 1, "local_constraint": {"house rule": null, "cuisine": null, "room type": null, "transportation": null, "event": null, "attraction": null}, "budget": 900, "level": "easy"}
```

QUERY:

Plan a 3-day trip for 1 person from Washington to Atlanta from November 18th to November 20th, 2024, with a budget of \$900.

—EXAMPLE 2—

JSON:

```
{"org": "Chicago", "dest": "Tennessee", "days": 5, "visiting_city_number": 2, "date": ["2024-11-02", "2024-11-03", "2024-11-04", "2024-11-05", "2024-11-06"], "people_number": 2, "local_constraint": {"house rule": null, "cuisine": null, "room type": "entire room", "event": null, "attraction": null, "transportation": null}, "budget": 2800, "level": "medium"}
```

QUERY:

Organize a 5-day itinerary for 2 people traveling from Chicago to explore 2 cities in Tennessee, between November 2nd and November 6th, 2024. The budget is \$2,800, and accommodations should include an entire room.

—EXAMPLE 3—

JSON:

```
{"org": "Tulsa", "dest": "California", "days": 7, "visiting_city_number": 3, "date": ["2024-11-01", "2024-11-02", "2024-11-03", "2024-11-04", "2024-11-05", "2024-11-06", "2024-11-07"], "people_number": 2, "local_constraint": {"house rule": null, "cuisine": null, "room type": "not shared room", "transportation": null, "event": ["Arts & Theatre", "Film"], "attraction": ["Museums", "Food & Drink"]}, "budget": 6000, "level": "hard"}
```

QUERY:

Create a detailed 7-day travel plan for 2 individuals starting from Tulsa and visiting 3 cities in California between November 1st and November 7th, 2024. The budget is \$6,000. Accommodations should be in a non-shared room. Include visits to museums and attractions involving food and drinks. The plan should also incorporate attending arts, theatre, and film events.

JSON:

C Case Studies

We showcase examples in this appendix that highlight the challenges within our dataset, underscoring its value in advancing LLMs' travel planning capabilities.

PoI List Time Exceeds the Departure Limit
Query: Plan a 3-day trip for 1 person from Denver to Santa Fe from November 1st to November 3rd, 2024, with a budget of \$1,200.
Plan: { ... "days": 3, "current_city": "from Santa Fe to Denver", "transportation": "Flight Number: F3932864, from Santa Fe to Denver, Departure Time: 13:05 , Arrival Time: 14:16", ... "point_of_interest_list": "Cozy cottage in central Santa Fe, stay from 07:00 to 08:30, nearest transit: Cerrillos @ 5th OB, 46.71m away; Tia Sophia's, visit from 09:00 to 09:50, nearest transit: Sandoval @ San Francisco OB, 104.85m away; Museum of International Folk Art, visit from 10:30 to 12:30, nearest transit: Cam. Lejo @ Museum of Int'l. Folk Art, 73.48m away; La Plazuela, visit from 13:00 to 14:00 , nearest transit: Cathedral @ Water, 130.45m away." }
Analysis: In this case, the plan recommends visiting an attraction after the departure time of the flight.

Timestamp Sequence Inconsistency
Query: Plan a 3-day trip for 3 people from Charleston to Washington from November 24th to November 26th, 2024, with a budget of \$3,750. Include visits to sights and landmarks.
Plan: { "days": 1, "current_city": "from Charleston to Washington", ... "point_of_interest_list": " King Bed Free Parking DCA Airport, stay from 09:30 to 08:30 , nearest transit: ENTRANCE B - FORT MYER DR., 360.84m away; Clyde's of Georgetown, visit from 09:40 to 10:30, nearest transit: ROSSLYN, 1120.37m away; White House, visit from 11:00 to 14:00, nearest transit: MCPHERSON SQ METRORAIL STATION - SW CORNER VERMONT AVE. & I ST., 409.99m away; Joe's Seafood Prime Steak & Stone Crab, visit from 14:20 to 15:20, nearest transit: MCPHERSON SQ, 150.59m away; The Capital Grille, visit from 20:45 to 22:00, nearest transit: ARCHIVES, 121.81m away; King Bed Free Parking DCA Airport, stay from 22:30 to 08:30, nearest transit: ENTRANCE B - FORT MYER DR., 360.84m away." ... }

Analysis:

In this case, the plan mis-structures the time sequence when mentioning the time duration to be spent at the accommodation.

Meal Timing Misalignment**Query:**

Plan a 3-day trip for 1 person from Birmingham to Detroit from November 19th to November 21st, 2024, with a budget of \$1,500.

Plan:

```
{ ...
"days": 3,
...
"dinner": "Vertical Detroit, Detroit",
...
"point_of_interest_list": "Modern Boutique Condo - "Au coeur de Detroit", stay
from 08:30 to 09:30, nearest transit: Sproat St - Southbound, 302.98m away;
Honest John's Bar & Grill - Selden St., visit from 09:40 to 10:30, nearest transit:
Mack Ave - Southbound, 509.79m away; The Ford Piquette Avenue Plant, visit
from 11:00 to 14:00, nearest transit: Amsterdam St - Southbound, 528.25m away;
Selden Standard, visit from 14:30 to 15:30, nearest transit: Canfield St
- Southbound, 586.42m away; Vertical Detroit, visit from 16:00 to 17:15 , nearest
transit: Grand Circus - Northbound, 257.87m away." }
```

Analysis:

In this case, the plan schedules dinner from 16:00 to 17:15, which is unrealistic and impractical.

Consistent PoI Time Duration Adherence**Query:**

Plan a 3-day trip for 1 person from Las Vegas to Fort Wayne from November 17th to November 19th, 2024, with a budget of \$2,100.

Plan:

```
{ "days": 1,
"current_city": "from Las Vegas to Fort Wayne",
"transportation": "Flight Number: F0373120, from Las Vegas to Fort Wayne,
Departure Time: 23:24, Arrival Time: 06:06 ",
"breakfast": "-",
"attraction": "-",
```



```

"lunch": "-",
"dinner": "-",
"accommodation": "Sunny Victorian 1-Bedroom apartment near Downtown,
Fort Wayne",
"event": "-",
"point_of_interest_list": "Sunny Victorian 1-Bedroom apartment near Downtown,
stay from 06:30 to 07:00 , nearest transit: Calhoun St. at Taber St. (SB), 76.02m
away." },
{ "days": 2,
"current_city": "Fort Wayne",
"transportation": "-",
"breakfast": "Sara's Family Restaurant, Fort Wayne",
"attraction": "Fort Wayne Museum of Art, Fort Wayne; Historic Fort Wayne,
Fort Wayne",
"lunch": "Fort Wayne's Famous Coney Island, Fort Wayne",
"dinner": "The Original Oley's Pizza, Fort Wayne",
"accommodation": "Sunny Victorian 1-Bedroom apartment near Downtown,
Fort Wayne",
"event": "-",
"point_of_interest_list": "Sunny Victorian 1-Bedroom apartment near Downtown,
stay from 07:00 to 08:30 , nearest transit: Calhoun St. at Taber St. (SB), 76.02m
away; Sara's Family Restaurant, visit from 09:00 to 09:50 , nearest transit: Kroger,
210.11m away; Fort Wayne Museum of Art, visit from 10:30 to 13:30 , nearest
transit: Main St. just past bus hut west of Lafayette St. (WB), 34.80m away;
Fort Wayne's Famous Coney Island, visit from 14:00 to 15:00 , nearest transit: W.
Main St. at Harrison St. (WB), 31.97m away; Historic Fort Wayne, visit from
15:30 to 18:30 , nearest transit: Spy Run Ave. at Baltes Ave. (NB), 92.88m away;
The Original Oley's Pizza, visit from 19:00 to 20:15 , nearest transit: Liberty Mills
Apts., 6101 Cornwallis Dr., 2376.66m away; Sunny Victorian 1-Bedroom
apartment near Downtown, stay from 21:00 to 07:00 , nearest transit: Calhoun St.
at Taber St. (SB), 76.02m away." },
{ "days": 3,
"current_city": "from Fort Wayne to Las Vegas",
"transportation": "Flight Number: F2558843, from Fort Wayne to Las Vegas,
Departure Time: 13:16 , Arrival Time: 14:01",
"breakfast": "Cosmos House Of Pancakes, Fort Wayne",
"attraction": "Fort Wayne Children's Zoo, Fort Wayne",
"lunch": "-",
"dinner": "-",
"accommodation": "-",
"event": "-",
"point_of_interest_list": "Sunny Victorian 1-Bedroom apartment near Downtown,
stay from 07:00 to 08:30 , nearest transit: Calhoun St. at Taber St. (SB), 76.02m
away; Cosmos House Of Pancakes, visit from 09:00 to 09:50 , nearest transit:
Canterbury Marketplace, 18.19m away; Fort Wayne Children's Zoo, visit from
10:30 to 12:30 , nearest transit: Franke Park, 101.97m away." }

```

Analysis:

In this case, the plan accurately follows predefined time constraints for Points of Interest (PoIs), ensuring structured scheduling.

Inconsistent Itinerary with Adventure Seeker Profile

Query:

Plan a 3-day trip for 2 people from Miami to Baltimore from November 3rd to November 5th, 2024, with a budget of \$1,800. The itinerary should include visits to zoos and aquariums, as well as fun and games attractions. Additionally, incorporate arts and theatre, and film events. Accommodations should allow pets.

Plan:

"persona": "Traveler Type: Adventure Seeker ; Purpose of Travel: Adventure; Spending Preference: Luxury Traveler; Location Preference: Mountains",

"plan": [

{ "days": 1,

...

"attraction": "National Aquarium, Baltimore" ,

"lunch": "Miss Shirley's Cafe, Inner Harbor",

"dinner": "Restaurante Tio Pepe",

"accommodation": "Charming Baltimore Home w/Parking - Dog Friendly!",

...

}

]

Analysis:

In this case, the plan fails to align with the Adventure Seeker persona type, which expects more than 2 attractions per day.

Constraint-Compliant Itinerary Generation

Query:

Plan a 3-day trip for 2 people from Portland to Baltimore from November 5th to November 7th, 2024, with a budget of \$1,500. The trip should include visits to Boat Tours & Water Sports and Sights & Landmarks. Note that children under 10 are allowed, and no self-driving transportation should be used.

Plan:

"persona": "Traveler Type: Laidback Traveler; Purpose of Travel: Relaxation; Spending Preference: Economical Traveler; Location Preference: Mountains",

"plan": [

{ "days": 1,

"current_city": "from Portland to Baltimore",

"transportation": "Flight Number: F0060549, from Portland to Baltimore,

Departure Time: 05:42, Arrival Time: 07:07",
 "breakfast": "Miss Shirley's Cafe, Inner Harbor, Baltimore",
 "attraction": "Top of the World Observation Level, Baltimore",
 "lunch": "Phillips Seafood, Baltimore",
 "dinner": "The Capital Grille, Baltimore",
 "accommodation": "Private Queen Suite (B) Fells and Hopkins Medical,
 Baltimore",
 "event": "-",
 "point_of_interest_list": "Private Queen Suite (B) Fells and Hopkins Medical, stay
 from 07:30 to 08:30, nearest transit: JOHNS HOPKINS STATION (METRO),
 821.12m away; Miss Shirley's Cafe, Inner Harbor, visit from 09:00 to 09:50,
 nearest transit: SHOT TOWER METRO West Entrance, 285.29m away; Top of the
 World Observation Level, visit from 10:30 to 13:30, nearest transit: SHOT TOWER
 METRO West Entrance, 453.53m away; Phillips Seafood, visit from 14:00 to
 15:00, nearest transit: SHOT TOWER METRO West Entrance, 344.26m away;
 The Capital Grille, visit from 19:00 to 20:15, nearest transit: SHOT TOWER
 METRO West Entrance, 293.21m away; Private Queen Suite (B) Fells and Hopkins
 Medical, stay from 21:00 to 07:00, nearest transit: JOHNS HOPKINS STATION
 (METRO), 821.12m away." },
 { "days": 2,
 "current_city": "Baltimore",
 "transportation": "-",
 "breakfast": "Blue Moon Cafe, Baltimore",
 "attraction": "Federal Hill Park, Baltimore",
 "lunch": "La Tavola, Baltimore",
 "dinner": "Restaurante Tio Pepe, Baltimore",
 "accommodation": "Private Queen Suite (B) Fells and Hopkins Medical,
 Baltimore",
 "event": "-",
 "point_of_interest_list": "Private Queen Suite (B) Fells and Hopkins Medical, stay
 from 07:00 to 08:30, nearest transit: JOHNS HOPKINS STATION (METRO),
 821.12m away; Blue Moon Cafe, visit from 09:00 to 09:50, nearest transit: SHOT
 TOWER STATION (METRO) sb, 1161.67m away; Federal Hill Park, visit from
 10:30 to 14:00, nearest transit: SHOT TOWER METRO West Entrance, 1089.59m
 away; La Tavola, visit from 14:30 to 15:30, nearest transit: SHOT TOWER
 STATION (METRO) sb, 472.69m away; Restaurante Tio Pepe, visit from 19:00 to
 20:15, nearest transit: LEXINGTON MARKET METRO North Entrance, 554.30m
 away; Private Queen Suite (B) Fells and Hopkins Medical, stay from 21:00 to
 07:00, nearest transit: JOHNS HOPKINS STATION (METRO), 821.12m away." },
 { "days": 3,
 "current_city": "from Baltimore to Portland",
 "transportation": "Flight Number: F0683635, from Baltimore to Portland,
 Departure Time: 21:41, Arrival Time: 23:01",
 "breakfast": "Miss Shirley's Cafe, Inner Harbor, Baltimore",
 "attraction": "Edgar Allan Poe's Grave Site and Memorial, Baltimore",

```

"lunch": "Dalesio's of Little Italy Restaurant, Baltimore",
"dinner": "Thames Street Oyster House, Baltimore",
"accommodation": "-",
"event": "-",
"point_of_interest_list": "Private Queen Suite (B) Fells and Hopkins Medical, stay
from 07:00 to 08:30, nearest transit: JOHNS HOPKINS STATION (METRO),
821.12m away; Miss Shirley's Cafe, Inner Harbor, visit from 09:00 to 09:50,
nearest transit: SHOT TOWER METRO West Entrance, 285.29m away; Edgar
Allan Poe's Grave Site and Memorial, visit from 10:30 to 13:30, nearest transit:
LEXINGTON MARKET (METRO) sb, 246.16m away; Dalesio's of Little Italy
Restaurant, visit from 14:00 to 15:00, nearest transit: SHOT TOWER STATION
(METRO) sb, 597.76m away; Thames Street Oyster House, visit from 19:00 to
20:15, nearest transit: SHOT TOWER STATION (METRO) sb, 1428.09m away." }
]

```

Analysis:

In this case, the plan successfully adheres to persona-based preferences and local constraints, ensuring alignment with user-defined travel expectations.

Unrealistic Transit Proximity Issue

Query:

Plan a 3-day trip for 1 person from Tampa to Bloomington from November 5th to November 7th, 2024, with a budget of \$1,650.

Plan:

```

{ "days": 1,
  "current_city": "from Tampa to Bloomington",
  ...
  "accommodation": "Arcade House - Hot Tub, Games, Grill, FirePit, FUN,
Bloomington",
  "event": "-",
  "point_of_interest_list": "Arcade House - Hot Tub, Games, Grill, FirePit, FUN,
stay from 07:00 to 08:30, nearest transit: Prow Rd @ Meadows Hospital,
251003.54m away ; Miller Park Zoo, visit from 09:00 to 12:30, nearest transit:
Belle & Park Square, 252594.59m away ; Baxter's American Grill, visit from
13:00 to 14:00, nearest transit: Prow Rd @ Meadows Hospital, 247021.85m away;
Janko's Little Zagreb, visit from 19:00 to 20:15, nearest transit: Kirkwood Ave @
B-Line Trail IB, 98.16m away; Arcade House - Hot Tub, Games, Grill, FirePit,
FUN, stay from 21:00 to 07:00, nearest transit: Prow Rd @ Meadows Hospital,
251003.54m away ." }

```

Analysis:

In this case, the plan selects PoIs whose nearest transit stop is hundreds of kilometers away from accommodations and attractions.

D Impact of Distribution Shift

To assess the impact of distributional variation, we partitioned the 3-day dataset into two approximately equal subsets, each annotated by a distinct group of annotators (roughly half per subset). As a result, the subsets reflect differing annotation distributions—effectively corresponding to different underlying parameter settings. The results presented in Table 19 are consistent with the trends discussed in Subsection 4.2. More generally, these findings highlight that model parameters can be adapted to accommodate the specific characteristics of any given annotation set, as described in Subsection 3.3.

Setting	Subset	\bar{T}_{meal}	\bar{T}_{attrac}	$\bar{S}_{spatial}$	$\bar{S}_{persona}$	\bar{S}_{ord}
<i>w/o Parameter Info</i>	I	0.491	0.0126	0.901	0.50	0.754
	II	0.590	0.0202	0.836	<u>0.49</u>	0.763
<i>w/ Parameter Info</i>	I	0.668	0.0119	0.903	0.51	0.772
	II	0.806	0.0308	0.837	<u>0.49</u>	0.768

Table 19: Impact of Distribution Shift: Results on subsets of the 3-day category

E Annotator Details

E.1 Guidelines for Annotators

The annotation process involves generating a travel plan that is both feasible and, if possible, optimal. The annotated plan must be based on reference information while considering constraints such as local preferences (e.g., cuisine type, attraction category) and traveler personas (e.g., laidback, economical). Additionally, common sense should be applied when selecting points of interest, and deviations from suggested durations or costs must be justified. A detailed breakdown of these annotation guidelines, including priority handling, public transit considerations, and documentation requirements, is provided in Table 20.

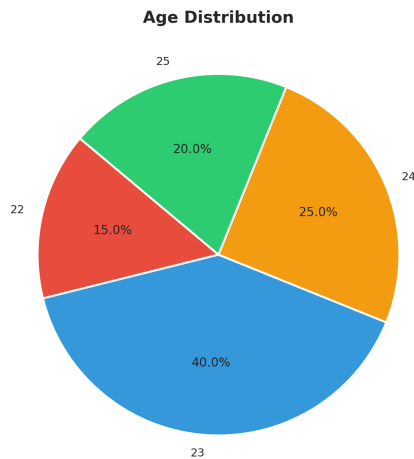


Figure 4: Age Distribution of our graduate student annotators.

E.2 Annotator Demographics

The annotator demographics, as illustrated by the figures, show a diverse range of experience levels and backgrounds. The years of English education vary significantly, with a distribution indicating that most annotators have between 12 to 20 years of formal English instruction (Figure 6). This suggests a predominantly well-educated group with proficiency in the language. Additionally, the age distribution skews toward early-to-mid adulthood, with a concentration around 22 to 25 years old (Figure 4), implying that most annotators are young graduate students. The gender distribution reflects participation from a diverse range of genders among the 25 graduate students. (Figure 5).

The combination of these factors suggests that the dataset is annotated by individuals with substantial English proficiency and cognitive maturity,

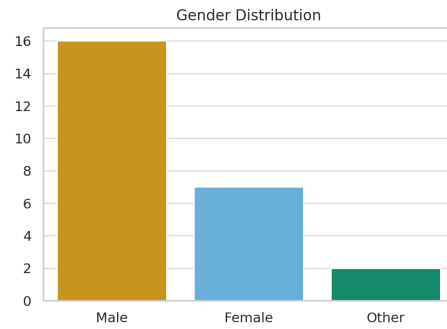


Figure 5: Gender Distribution of our graduate student annotators.

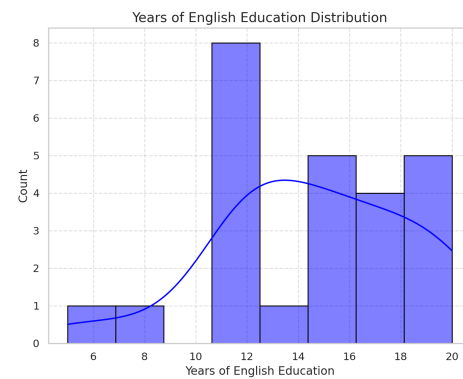


Figure 6: Years of Formal English Education Statistics of our graduate student annotators.

contributing to reliable and contextually aware annotations.

#	Annotation Guideline
1	The goal of annotation is to generate a feasible travel plan that meets the query requirements. If multiple valid plans exist, selecting the most optimal plan is encouraged.
2	Every element in the annotated travel plan must be grounded in reference data relevant to the given query. No fabricated or unverifiable information should be included.
3	Common sense should be maintained when selecting travel plans. Refer Table 11.
4	Any local constraints specified in the query must be respected. These may include preferences for cuisine types (e.g., Indian, Mediterranean) or attraction categories (e.g., Sights & Landmarks, Zoos & Aquariums).
5	The travel plan should align with the traveler’s persona. For instance, a laidback traveler would prefer a schedule with 1-2 attractions per day, even if more options exist. An economical traveler would favor budget-friendly choices over expensive alternatives.
6	If there is a conflict between local constraints and traveler persona preferences, the local constraints must be prioritized. If no valid plan can be formed while satisfying local constraints, a justification must be provided in Remarks.
7	When selecting Points of Interest (PoIs), priority should be given to those with a public transit stop within 5km. If choosing between a PoI that meets local constraints but lacks transit access and one that is transit-friendly but does not meet constraints, the former should be preferred—unless a better alternative exists.
8	The values for visit duration at attractions and average cost at restaurants should be treated as reference values rather than strict limits. Annotators are allowed to adjust these values if necessary but must document any deviations in the Remarks section.
9	Annotators should use their best judgment to ensure that the generated plans are practical and reasonable. Any significant decision-making considerations should be explicitly noted in the Remarks section.

Table 20: Guidelines for Annotation of Travel Plans and Remarks