

# ChinaTravel: An Open-Ended Benchmark for Language Agents in Chinese Travel Planning

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Recent advances in LLMs have spurred the development of *Language Agents* for real-world applications such as travel planning, which involves complex multi-constraint challenges. Existing benchmarks, however, often oversimplify reality with synthetic queries and limited constraints. To bridge this gap, we introduce *ChinaTravel*, the first open-ended benchmark based on authentic travel needs. We develop a domain-specific language (DSL) for compositional evaluation covering feasibility, constraints, and preferences. Experiments show neuro-symbolic agents achieve a 37.0% constraint satisfaction rate on human queries, a 10× improvement over neural models, demonstrating their potential in complex planning scenarios.

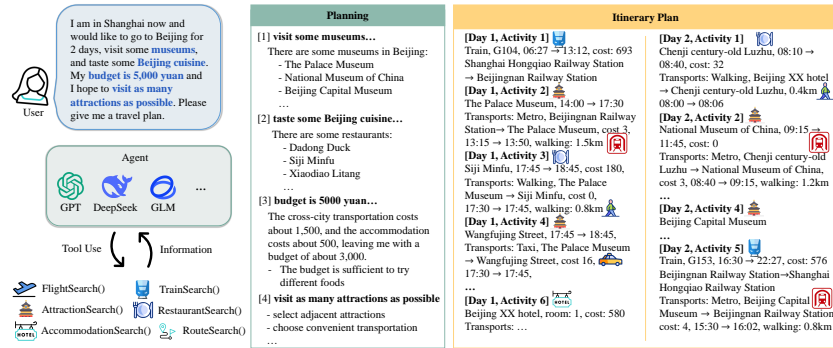


Figure 1: **Overview of ChinaTravel.** Given a query, language agents employ various tools to gather information and plan a multi-day multi-POI itinerary. The agents are expected to provide a feasible and reasonable plan while satisfying the **logical constraints** and **preference requirements**.

## 1 Introduction

A long-standing goal in AI is to build reliable planning agents capable of assisting humans in real-world tasks. Among numerous tasks [19, 27, 23, 11], travel planning stands out as a significant domain, presenting academic challenges and practical value due to its inherent complexity. It requires integrating information from various tools and making interdependent spatial, temporal, and financial decisions while satisfying user requirements. To evaluate language agents on it, Xie et al. [24] introduced the TravelPlanner benchmark, which suffers from two key limitations: (1) U.S.-centric intercity bias instead of common multi-day city trips; (2) synthetic and templated queries rather than diverse human needs. Shortly after its release, Hao et al. [10] achieved a 97% success rate using a neuro-symbolic method with formal verification, highlighting TravelPlanner’s oversimplification.

To address the gap, we introduce ChinaTravel, an open-ended benchmark for multi-POI travel planning (Fig. 1) with authentic Chinese queries and compositional constraint evaluation. It offers a more realistic and challenging benchmark for real-world travel planning. Key contributions include: (1)

Table 1: ChinaTravel’s Domain-Specific Language (DSL) for logical constraints.

Name	Syntax	Description
variables	$x, y, z, \dots$	Variables that refer to activities in the travel planning domain.
not	$\text{not } \text{expr}$	The negation of an Boolean-valued expression.
and, or	$\text{expr}_1 \text{ and } \text{expr}_2$	The conjunction/disjunction of an Boolean-valued expression.
<, >, ==	$\text{expr}_1 < \text{expr}_2$	Return an expression with built-in number comparison functions.
+, -, *, /	$\text{expr}_1 + \text{expr}_2$	Return an expression with built-in number calculation functions.
attributes	$\text{cost}(\text{var})$	A function that takes activities as inputs and returns the attributes, such as cost, type or time.
relation	$\text{dist}(\text{expr}_1, \text{expr}_2)$	A function that takes locations as inputs and returns the distance.
effect	$\text{var} = \text{expr}$	An assignment affects a variable $\text{var}$ with the expression $\text{expr}$ .
union, inter, diff	$\text{uni}(\{\text{var}\}_1, \{\text{var}\}_2)$	Return a set with the built-in union/intersection/difference operations of given two sets.
enumerate	$\text{for } \text{var} \text{ in } \{\text{var}\}$	Enumerate all variables in the collection $\{\text{var}\}$ .
when	$\text{if } \text{expr} : \text{effect}$	The conditional effect takes a Boolean-valued condition of the expression $\text{expr}$ , and the effect $\text{effect}$ .

23 **Comprehensive Evaluation Framework:** Includes a DSL for scalable requirement formulation and  
24 automated evaluation, and diverse metrics for feasibility, constraints, and preferences. (2) **Integration**  
25 **of Synthetic and Human Data:** Beyond LLM-generated queries, our validation set contains 154  
26 human queries with complex constraints, while the test set has 1,000 open scenarios, assessing  
27 generalization to unseen constraints. (3) **Empirical Neuro-Symbolic Insights:** Experiments show  
28 neuro-symbolic agents significantly outperform neural methods, achieving 27.9% constraint satisfac-  
29 tion vs. 2.60%. (4) **Identified Challenges for Future Research:** Open contextual reasoning and  
30 unseen concept composition, guiding future research toward real-world applicability.

## 31 2 ChinaTravel Benchmark

32 Motivated by China’s substantial travel demand, ChinaTravel provides a sandbox environment for  
33 generating multi-day itineraries with multiple POIs across specified cities. It is meticulously designed  
34 to provide a comprehensive and scalable evaluation framework in travel planning, encompassing three  
35 critical dimensions: environmental feasibility, constraint satisfaction, and preference comparison.

36 **Environment Information.** ChinaTravel provides a sandbox environment with real-world travel  
37 data from 10 popular Chinese cities. It includes 720 flights and 5770 trains with detailed schedules  
38 and prices, as well as 3413 attractions, 4655 restaurants, and 4124 hotels, each annotated with name,  
39 location, opening hours, price, and type. It simulates real-market APIs for realistic querying. We  
40 impose 25 constraints across dietary, accommodation, transportation, temporal, spatial, and attraction  
41 domains to ensure plan feasibility (e.g., POI existence and transport validity). Details are in App. D.1.

42 **Logical Constraint.** A key capability in travel planning is satisfying personalized user needs. We  
43 extend the logical constraint framework from TravelPlanner [24] by introducing a Domain-Specific  
44 Language (DSL) for general compositional reasoning. ChinaTravel’s DSL comprises pre-defined  
45 concept functions (Tab. 1), enabling flexible constraint representation. Unlike TravelPlanner, which  
46 relies on only five fixed concepts (e.g., total budget, room rules), our approach supports open-ended  
47 logical requirements, such limiting dining costs to 1000 CNY or ensuring arrival in Shanghai before  
48 6 PM on day two, without manual intervention. By composing concepts in Python-like syntax  
49 (Fig. 5a, 5b), the DSL facilitates automated plan validation via a Python compiler, significantly  
50 enhancing evaluation capability. Further examples and a tutorial are provided in App. D.2.

51 **Preference Requirement.** In addition to hard constraints, travel planning must accommodate  
52 soft preferences, which involve quantitative comparisons of continuous values rather than Boolean  
53 satisfaction. Common examples include maximizing the number of attractions, minimizing transit  
54 time, or prioritizing nearby POIs. We formalize them as minimization or maximization objectives via  
55 our DSL, enabling automated evaluation. An example for maximizing attractions is shown in Fig. 5c.

56 **Benchmark Construction.** ChinaTravel constructs user queries via a four-stage pipeline: (I) Manual  
57 design of database and APIs using collected multi-day travel information and commercial-like API  
58 interfaces (App. D.1). (II) LLM-based generation of query skeletons with natural language variation,  
59 stratified by complexity (e.g., *Easy* with one constraint; see App. D.3). (III) Quality control via  
60 manual checks and automated validation of constraint satisfaction using DSL-executable heuristic  
61 search. (IV) Collection of human requirements: 154 queries form the Human-154 set; 1000 via  
62 survey platform WJX comprise Human-1000, both with DSL-based auto-evaluation.

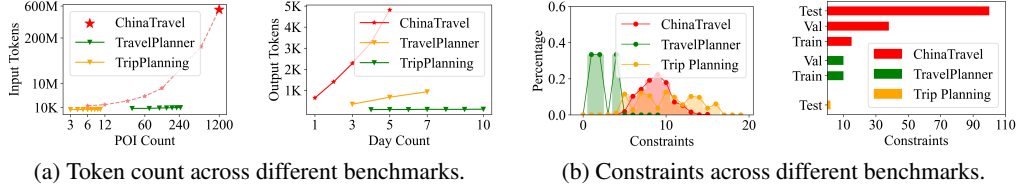


Figure 2: (a) ChinaTravel’s fine-grained spatiotemporal planning demands extremely larger input/output text volumes than existing benchmarks, posing fundamental challenges to text-wise planning. (b) ChinaTravel’s authentic requirements, with combinatorial scalable constraints formulation, systematically surpasses conventional closed-form benchmarks in diversity and openness.

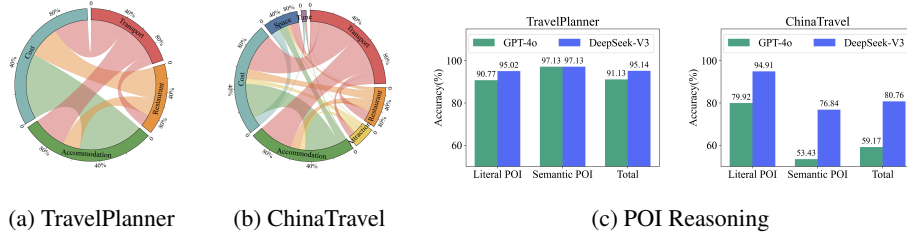


Figure 3: Co-occurrence distribution of different constraints on TravelPlanner (a) and ChinaTravel’s Human1000 (b). (c) The unsatisfactory performance of advanced LLMs on the auxiliary task, POI reasoning, reveals the challenges of open contextual reasoning in ChinaTravel’s dataset.

























### 3 Benchmark Characteristics

**Context-Rich Long-Horizon Planning.** ChinaTravel poses unprecedented contextual complexity compared to existing benchmarks, NaturalPlan [30] and TravelPlanner [24]. As quantified in Fig. 2a, (1) Processing over 1200 candidate POIs per query (4× TravelPlanner max, 120× Trip Planning). (2) Generating 540M contextual tokens from dense POI networks, surpassing both DeepSeek-V3 (64K) and GPT-4o (128K) capacities, even aggressive 6-POI downsampling retains 40K tokens (Fig. 2a). (3) Requiring 4.8K output tokens for 5-day plans, versus 0.9K (TravelPlanner’s 7-day) and 0.5K (Trip Planning’s 30-day) [28]. These demands reveal the inadequacy of single-pass text generation. Effective solutions may require hierarchical decomposition or symbolic planning, with iterative subtask execution for scalable long-horizon reasoning [28].






**Diversity and Openness of Travel Requirements.** ChinaTravel models more diverse requirements than TravelPlanner and NaturalPlan. Fig. 2b shows: (1) It exhibits a Gaussian-style constraint distribution (6–12 per query), contrasting with TravelPlanner’s simplicity ( $\leq 5$ ) and TripPlanning’s limited types. (2) It supports exponential constraint combinations, from 15 (synthetic) to 100 (human1000 test), including 88 novel types via DSL composition (Tab. 1). (3) Constraint co-occurrence follows a long-tailed Zipf distribution [1] (Fig. 3b), unlike TravelPlanner’s uniform pattern. For example, cost constraints strongly correlate with transportation and accommodation, reflecting real-world travel behavior. These features, derived from open-ended user studies, present complex composite reasoning challenges, verifiable under our framework (Sec. 3.2).

**Open Contextual Reasoning.** Travel requirements often involve contextual ambiguity not directly mapped to database attributes. For example, “local cuisine” refers to Benbang in Shanghai versus Beijing cuisine in Beijing, and “no spicy food for children” implies excluding Sichuan and Chongqing cuisines. These observations arise the necessity for real-world travel agents to conduct open contextual reasoning that bridges arbitrary user expressions with verifiable symbolic semantics in databases, a evaluation capability inadequately supported by existing synthetic benchmarks like TravelPlanner. To study this, we design a POI reasoning task: replacing POIs in DSL constraints with  $\langle$ placeholder $\rangle$  tags, requiring LLMs to masked-DSL sentences through contextual reasoning. POIs are categorized as Literal (explicitly mentioned in user queries) or Semantic (requiring cultural/contextual inference). 78.4% of DSL statements in Human1000 contain Semantic POIs versus only 5.4% in TravelPlanner. Both models achieve the accuracy over 90% on TravelPlanner, where semantic POIs follow simplistic synthetic patterns. However, on ChinaTravel’s authentic Semantic POIs, performance significantly declines (DS: 94%  $\rightarrow$  76%, GPT: 79%  $\rightarrow$  53%, Fig. 3c). This gap highlights the critical challenge of real-world contextual understanding in travel planning.


Table 2: Main results of different LLMs and planning strategies on the ChinaTravel benchmark.

		DR	EPR		LPR		C-LPR		FPR	DR	EPR		LPR		C-LPR		FPR
			Mic.	Mac.	Mic.	Mac.					Mic.	Mac.	Mic.	Mac.			
		Easy (#300)								Human-Val (#154)							
Act		70.4	49.9	0	64.6	30.8	0	0			-						
		<b>97.5</b>	70.8	0	86.8	68.8	0	0			-						
ReAct (zero-shot)		43.3	40.8	0	41.9	19.6	0	0		36.4	29.5	0.65	35.2	16.2	0.38	0	
		95.4	48.2	0	71.3	32.9	0	0		<b>96.1</b>	50.5	0	<b>72.4</b>	32.5	0	0	
ReAct (one-shot)		77.5	68.3	6.25	74.1	52.5	5.77	5.42		55.2	<b>57.3</b>	2.60	64.6	44.2	1.71	2.60	
		94.2	68.1	0	<b>89.4</b>	<b>70.8</b>	0	0		69.5	46.3	0	63.6	46.8	0	0	
NeSy Planning		75.3	<b>75.3</b>	75.3	70.4	52.6	70.4	52.6		51.9	53.2	52.5	47.0	<b>37.6</b>	46.5	<b>37.0</b>	
		75.0	73.6	<b>64.0</b>	73.5	63.3	<b>61.7</b>	<b>60.6</b>		45.4	50.1	45.4	40.9	29.8	<b>38.5</b>	27.9	
		72.3	67.0	34.0	70.4	49.6	32.6	28.3		42.8	47.4	42.2	36.2	27.2	34.4	25.3	
		32.0	31.9	31.3	29.1	21.0	28.3	21.0		25.9	25.8	24.0	22.3	12.3	20.5	11.0	
		30.3	30.3	30.3	27.6	19.6	27.6	19.6		37.6	38.2	37.6	32.7	18.8	32.2	18.8	
TTG (oracle)		18.3	21.5	8.66	17.2	15.0	8.23	8.66		9.09	12.8	2.59	7.65	5.19	2.39	1.29	
LLM-Modulo* (Oracle Verifier)		48.3	94.5	4.33	58.4	43.6	4.11	4.33		61.6	90.2	2.59	75.9	51.2	2.75	2.59	
		91.6	88.2	7.66	<b>95.5</b>	<b>84.6</b>	7.66	7.00		91.5	87.2	3.24	<b>92.9</b>	<b>66.2</b>	2.87	3.24	
		30.0	80.5	0.0	62.7	25.0	0.0	0.0		35.0	75.3	0.0	61.6	19.4	0.0	0.0	
		28.6	69.4	0.0	55.2	8.33	0.0	0.0		19.4	74.1	0.0	43.4	5.19	0.0	0.0	
		10.3	90.5	0.0	39.1	9.0	0.0	0.0		3.24	<b>92.2</b>	0.0	31.4	4.54	0.0	0.0	
NeSy Planning* (Oracle Translation)		<b>82.6</b>	<b>81.7</b>	<b>75.0</b>	<b>82.2</b>	75.3	<b>75.0</b>	<b>74.0</b>		<b>58.4</b>	59.6	<b>57.7</b>	53.8	46.1	<b>52.0</b>	<b>45.4</b>	
		66.6	66.7	66.0	64.6	63.6	64.6	62.6		52.6	46.9	42.9	47.6	40.9	43.9	40.9	
		69.3	69.3	59.3	70.2	59.6	59.3	57.9		53.2	55.1	54.5	48.0	42.8	47.6	40.9	
		52.6	52.6	52.6	50.4	45.3	50.4	45.6		40.9	42.8	42.8	37.7	28.5	37.7	27.9	
		33.3	33.2	32.6	32.1	32.0	31.4	32.3		29.2	29.1	26.6	25.4	20.1	23.4	19.4	
		Human-Test (#1000)								NeSy Planning* (Oracle Translation)							
NeSy Planning		<b>44.6</b>	<b>44.5</b>	<b>42.6</b>	<b>38.7</b>	<b>23.3</b>	<b>37.6</b>	<b>23.3</b>		<b>60.6</b>	<b>60.3</b>	<b>59.0</b>	<b>53.6</b>	<b>32.0</b>	<b>52.5</b>	<b>31.6</b>	
		37.3	37.2	35.0	30.7	11.3	29.2	11.3		27.8	27.8	27.1	24.8	12.8	24.4	12.8	
		36.6	36.5	34.6	29.6	6.43	28.5	6.43		41.1	41.1	40.6	34.6	13.8	34.2	13.8	

## 4 Empirical Study

**LLMs.** We evaluate the state-of-the-art LLMs,  DeepSeek-V3,  OpenAI GPT-4o, recognized for their world-leading performance. We also examine the open-source LLMs,  Qwen3-8B,  Llama3.1-8B, and  Mistral-7B, selected based on their computationally efficient 7B/8B architectures.

**Methods.** Pure-LLM-based ReAct [27], and its Act-only variant. Neuro-symbolic: TTG [12], which converts natural language needs into mixed-integer linear programming. LLM-modulo [13, 7], employing ground-truth symbolic verification to guide iterative LLM self-refinement. NeSy Planning (App. F), extending prior NeSy pipelines [10, 17, 26, 25] through our DSL enhancements.

**Main Results.** While pure LLMs generate structurally sound plans (high DR), they perform poorly in satisfying constraints (near-zero EPR/FPR). Our NeSy Planning effectively mitigates these limitations through neural-symbolic integration, significantly improving constraint satisfaction, achieving FPRs (52.6% on easy, 23.3% on human-1000, ), demonstrating robust generalization in challenging, constraint-rich scenarios. More deeper analysis is provided in the App. H.

## 5 Conclusion

We present ChinaTravel, a benchmark for multi-day multi-POI travel planning focused on authentic Chinese needs. We address the limitations of previous benchmarks by incorporating open-ended and diverse human queries, capturing real-world user needs. Additionally, we propose a scalable evaluation framework based on DSL, enabling comprehensive assessments of feasibility, constraint satisfaction, and preference comparison. These advancements provide a foundation for developing language agents capable of meeting diverse user requirements and delivering reliable travel solutions.



## References

- [1] Lada A Adamic and Bernardo A Huberman. Zipf’s law and the internet. *Glottometrics*, 3(1): 143–150, 2002.
- [2] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.
- [3] Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. TravelAgent: An AI assistant for personalized travel planning. *arXiv preprint arXiv:2409.08069*, 2024.
- [4] Wang-Zhou Dai, Qiu-Ling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. In *Advances in Neural Information Processing Systems*, pages 2811–2822, 2019.
- [5] Shujie Deng, Honghua Dong, and Xujie Si. Enhancing and evaluating logical reasoning abilities of large language models. In *Proceedings of the ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.
- [6] Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, et al. Miplib 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.
- [7] Atharva Gundawar, Mudit Verma, Lin Guan, Karthik Valmeekam, Siddhant Bhambri, and Subbarao Kambhampati. Robust planning with llm-modulo framework: Case study in travel planning. *arXiv preprint arXiv:2405.20625*, 2024.
- [8] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.
- [9] Sajal Halder, Kwan Hui Lim, Jeffrey Chan, and Xiuzhen Zhang. A survey on personalized itinerary recommendation: From optimisation to deep learning. *Applied Soft Computing*, 152: 111200, 2024.
- [10] Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. Large language models can solve real-world planning rigorously with formal verification tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, Albuquerque, New Mexico, 2025.
- [11] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- [12] Da Ju, Song Jiang, Andrew Cohen, Aaron Foss, Sasha Mitts, Arman Zharmagambetov, Brandon Amos, Xian Li, Justine Kao, Maryam Fazel-Zarandi, et al. To the globe (ttg): Towards language-driven guaranteed travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 240–249, 2024.
- [13] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: Llms can’t plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, Vienna, Austria, 2024.
- [14] Weiyu Liu, Geng Chen, Joy Hsu, Jiayuan Mao, and Jiajun Wu. Learning planning abstractions from language. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- [15] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepprolog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, pages 3753–3763, 2018.

- [16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing Atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.
- [17] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3806–3824, 2023.
- [18] Vibhor Sharma, Monika Goyal, and Drishti Malik. An intelligent behaviour shown by chatbot system. *International Journal of New Technology and Research*, 3(4):263312, 2017.
- [19] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.
- [20] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.
- [21] Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Kebin Hou, Dingyi Zhuang, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, and Wei Ma. Synergizing spatial optimization with large language models for open-domain urban itinerary planning. *CoRR*, abs/2402.07204, 2024.
- [22] Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6545–6554, 2019.
- [23] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864, 2023.
- [24] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [25] Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. Deliberate reasoning for llms as structure-aware planning with accurate world model. *CoRR*, abs/2410.03136, 2024.
- [26] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.
- [28] Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. Longproc: Benchmarking long-context language models on long procedural generation. *arXiv preprint arXiv:2501.05414*, 2025.
- [29] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. Huatuoogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 10859–10885, 2023.
- [30] Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V Le, Ed H Chi, et al. Natural plan: Benchmarking llms on natural language planning. *arXiv preprint arXiv:2406.04520*, 2024.

214	<b>Contents</b>	
215	<b>1 Introduction</b>	<b>1</b>
216	<b>2 ChinaTravel Benchmark</b>	<b>2</b>
217	<b>3 Benchmark Characteristics</b>	<b>3</b>
218	<b>4 Empirical Study</b>	<b>4</b>
219	<b>5 Conclusion</b>	<b>4</b>
220	<b>A Limitations</b>	<b>8</b>
221	<b>B Broader impacts</b>	<b>8</b>
222	<b>C Discussion with Related Work</b>	<b>8</b>
223	<b>D Detailed Design of ChinaTravel</b>	<b>9</b>
224	D.1 Sandbox Information . . . . .	9
225	D.2 Tutorial of DSL Expression . . . . .	9
226	D.3 Query Synthesis . . . . .	13
227	D.4 Data Diversity and Bias Mitigation . . . . .	19
228	D.5 Data with Preference . . . . .	19
229	D.6 Benchmark Difficulty and Applicability . . . . .	20
230	<b>E Discussion with TravelPlanner</b>	<b>20</b>
231	<b>F NeSy Planning</b>	<b>21</b>
232	<b>G Evaluation Metric in Competition</b>	<b>22</b>
233	<b>H Detailed Empirical Analysis</b>	<b>23</b>
234	H.1 Main Results . . . . .	23
235	H.2 Ablation Study with Preference . . . . .	25
236	<b>I Additional Experimental Results</b>	<b>25</b>
237	I.1 Multi-Preference Comparison . . . . .	25
238	I.2 Open Reasoning with Chinese Context . . . . .	26
239	I.3 Analysis of Pure-LLM Methods . . . . .	27
240	<b>J Statements about Scientific Artifacts</b>	<b>28</b>
241	<b>K Statements about Human Participants</b>	<b>28</b>
242	K.1 Instructions Given To Participants . . . . .	28
243	K.2 Recruitment And Payment . . . . .	28

244	K.3 Data Consent . . . . .	29
245	K.4 Characteristics of Annotators . . . . .	29
246	K.5 DSL Annotation for Human Data . . . . .	29
247	<b>L TTG</b>	<b>30</b>
248	L.1 Constraints Formulation . . . . .	30
249	L.2 Experiment . . . . .	30

## 250 **A Limitations**

251 Our research represents a significant step forward in evaluating the travel planning capabilities of  
252 language agents, but it is not without challenges. One limitation lies in its focus on Chinese travel  
253 planning. Due to the inherent differences in natural language, the translated versions of queries may  
254 fail to fully capture the challenges of understanding requirements in Chinese queries, potentially  
255 limiting its applicability in a global context. However, given the substantial demand within China’s  
256 travel market, we believe a benchmark tailored to Chinese travel planning is both necessary and  
257 socially valuable. Although our benchmark is comprehensive, it may not encompass the full range  
258 of requirements encountered in real-world scenarios. The high cost of collecting authentic data  
259 has limited the number of human queries in our study. To address this, future work will focus on  
260 combining LLMs with real user queries to automate the generation of a wider variety of human-like  
261 queries. Continuous refinement and expansion of our benchmark are crucial for more accurately  
262 reflecting the realistic travel planning needs.

## 263 **B Broader impacts**

264 The ChinaTravel benchmark primarily serves as a foundational research tool to evaluate language  
265 agents in complex, real-world travel planning scenarios. By providing an open-ended benchmark  
266 grounded in authentic Chinese travel requirements, this work aims to advance the development of  
267 reliable and generalizable AI systems for practical planning tasks. Its positive societal impacts include:  
268 (1) Improved Travel Planning Efficiency: By rigorously testing agents’ ability to handle multi-day  
269 itineraries and combinatorial constraints, this benchmark encourages the creation of more robust AI  
270 assistants, potentially reducing the time and effort users spend on organizing trips. (2) Validation for  
271 Real-World Applications: The benchmark establishes a critical foundation for deploying language  
272 agents in practical travel planning settings, where multi-objective planning and constraint satisfaction  
273 are essential. (3) Promotion of Open Research: The release of this benchmark bridges cutting-edge  
274 LLMs with classical neuro-symbolic AI paradigms, fostering cross-disciplinary collaboration between  
275 academia and industry. It promotes the reliable, constraint-aware AI systems, while accelerating  
276 innovation in both foundational planning capabilities and real-world deployment scenarios.

277 Potential negative impacts largely depend on how future systems built upon this benchmark are  
278 deployed. For instance: (1) Bias and Fairness: If agents inherit biases from training data or misalign  
279 with diverse user preferences, they might disproportionately recommend certain POIs or services.  
280 Mitigation requires ongoing fairness audits and inclusive data practices. (2) Misuse Risks: Malicious  
281 actors could exploit highly capable planning agents to generate deceptive itineraries or manipulate  
282 travel services. Such risks underscore the need for ethical guidelines and safeguards in downstream  
283 applications. As a benchmark, ChinaTravel itself does not directly deploy agents but focuses on  
284 evaluation. Its design emphasizes transparency, verifiability, and scalability, aligning with broader  
285 efforts to ensure AI systems are both effective and controllable. Future work should prioritize  
286 responsible use, including robust validation of real-world systems and addressing socio-technical  
287 challenges like bias mitigation and user privacy.

## 288 **C Discussion with Related Work**

289 **LLM-based Agents** have demonstrated significant capability in understanding complex instructions  
290 and employing domain-specific tools to complete tasks, showcasing their potential in fields such

as visual reasoning [8], healthcare [29] and robotics [14]. This reduces the reliance of previous agents on domain-specific efforts, that is, either mainly following domain-specific rules to plan (rule-based agents, such as DeepBlue [2] and Eliza [18]) or mainly learning from domain-specific data to plan (reinforcement-learning-based agents, such as AlphaGo [20] and Atari DQN [16]). While the language agents have shown promising results in some domains, most of their planning scenarios are limited to simple tasks with single objective function and fail in the travel planning benchmark with complex logical constraints.

**Neuro-Symbolic Learning** explores to combine traditional symbolic reasoning with learning to enhance the reliability [15, 22, 4]. In the era of large language models, Pan et al. [17] presents the LogicLM integrates LLMs with separate symbolic solvers for various logical reasoning tasks. They first utilize LLMs to translate a natural language problem into a symbolic formulation. Afterward, a deterministic symbolic solver performs inference on the formulated problem to ensure the correctness of the results. Deng et al. [5] supplement LogicLM with a Self-Refinement Module to enhance the reliability of LLM translation. In the travel planning domain, Hao et al. [10] presents a framework with a similar pipeline. It first extracts the logical constraints from natural language queries and then formalizes them into SMT code. Thanks to SMT solvers being sound and complete, this neuro-symbolic solution guarantees the generated plans are correct and has basically solved the TravelPlanner benchmark with a 97% pass rate.

**Travel Planning** is a time-consuming task even for humans, encompassing travel-related information gathering, POI selection, route mapping, and customization to meet diverse user needs [9]. Natural languages are one of the most common ways for users to express their travel requirements. However, the ambiguity and complexity of travel requirements make it still challenging for LLMs to generate accurate and reliable travel plans. Xie et al. [24] presents the TravelPlanner benchmark for cross-city travel planning and reveals the inadequacies of pure-LLM-driven agents. TravelPlanner generates user queries through LLMs and provides a rigorous evaluation mechanism to verify whether the provided plans can meet the logical constraints in the queries. It has become a pivotal benchmark for language agents in real-world travel planning. Tang et al. [21] study the open-domain urban itinerary planning where a single-day multi-POI plan is required. They integrate spatial optimization with large language models and present a system ITTNERA, to provide customized urban itineraries based on user needs. A concurrent work, TravelAgent [3], also considers a multi-day multi-POI travel planning problem for the specified city. It constructs an LLM-powered system to provide personalized plans. However, due to the high cost of collecting and annotating real travel needs, they evaluate the proposed TravelAgent in only 20 queries. This also demonstrates the necessity of introducing a new benchmark for travel planning.

## D Detailed Design of ChinaTravel

### D.1 Sandbox Information

We started collecting travel information with the motivation of planning a multi-day, multi-POI itinerary in four aspects: attractions, accommodation, activities, and transportation. Developers first determine the POI description information that needs to be obtained from the user’s perspective, such as cuisine and hotel features. Based on this feature set, we collect public information to construct the database. For the design of APIs, we directly support queries based on the regular expressions from agents. At the same time, we expect the design of APIs to have similar features and characteristics to existing commercial APIs, enabling our dataset to be applicable to more realistic scenarios. The information our database contains is shown in Table 3 and the APIs we offer is in Table 4. Table 5 shows the information of environment constraints in ChinaTravel.

### D.2 Tutorial of DSL Expression

Here is a tutorial, that provides a step-by-step guide to utilizing ChinaTravel’s Domain-Specific Language (DSL) with predefined concept functions for expressing logical constraints and preferences.

**DSL Overview** In the main body of this paper, we have detailed the basics of our DSL in the Table 1. The DSL is a Python-like language designed to formalize travel planning requirements into executable code. It enables automated validation of itineraries against user constraints and preferences. Key components include: 1) *Concept Functions*: Predefined functions (e.g., `activity_cost`, `poi_distance`)

Tool	Information
Attractions	Name, Type, Latitude, Longitude, Opentime, Endtime, Price, Recommendmintime, Recommendmaxtime
Accommodations	Name, Name_en, Featurehoteltype, Latitude, Longitude, Price, Numbed
Restaurants	Name, Latitude, Longitude, Price, Cuisinetype, Opentime, Endtime, Recommendedfood
Transportation	Transportation in specific city including walk, metro and taxi
IntercityTransport	Flight: FlightID, From, To, BeginTime, EndTime, Duration, Cost Train: TrainID, TrainType, From, To, BeginTime, EndTime, Duration, Cost
Poi	Names of POIs(including intercity transportation hub) and their coordinates

Table 3: Sandbox Information

343 that extract attributes from travel plans. 2) *Operators*: Logical (and, or, not), arithmetic (+, -, \*, /),  
344 and comparison operators (<, >, ==). 3) *Control Structures*: Loops (for), conditionals (if), and set  
345 operations (union, intersection). More examples are provided in Fig. 4.

DSL Syntax Compliance	Open Language Reasoning	Unseen Concept Composition
<b>Query:</b> Four of us would like to visit Shanghai for 2 days, try local fast food, and <b>stroll along the Bund</b> . Please provide us with a travel plan. <b>DSL Translation (Qwen2.5-7B):</b> <code>result=(activity_position(activity)=='外滩') specified</code> <b>Error:</b> name 'activity' is not defined invalid syntax (<string>) 	<b>Query:</b> I am currently in Nanjing and would like to go on a 5-day trip to Beijing with a friend. We plan to travel by high-speed train both ways and hope to <b>try some local specialty foods</b> . <b>DSL Translation (GPT-4o):</b> <pre>restaurant_type_set = set() for activity in allactivities(plan):     if activity_type(activity) in ['breakfast', 'lunch', 'dinner']:         restaurant_type_set.add(restaurant_type(activity, target_city(plan))) result=('本帮菜')&lt;=restaurant_type_set</pre>	<b>Query:</b> I am traveling alone from Nanjing to Shanghai in the morning for a day trip. I plan to visit a university campus and return in the evening, <b>making sure to catch the train back before 7 PM</b> . <b>DSL Translation (GPT-4o):</b> <pre>result = True for activity in allactivities(plan):     if activity_end_time(activity) &gt;= '19:00':         result = False</pre>
<b>Query:</b> Current location: Guangzhou. I want to go to Shenzhen alone for a day, with a budget of 1000 RMB. Please provide me with a travel plan. <b>DSL Translation (Qwen2.5-7B):</b> <pre>result=True for activity in allactivities(plan):     if activity_type(activity) in ['train', 'airplane']:         intercity_transport_set.add(activity_type(activity)) result=(intercity_transport_set=='train')</pre> <b>Error:</b> name 'intercity_transport_set' is not defined 	<b>Query:</b> My parents and I plan a five-day travel from Nanjing to Beijing to watch the flag-raising ceremony, and <b>we want to stay at a hotel near Tiananmen Square</b> . <b>DSL Translation (GPT-4o):</b> <pre>hotel_names_set = set() for activity in allactivities(plan):     if activity_type(activity)=='accommodation':         hotel_names_set.add(activity_position(activity)) result=('秋果S1918庭院式酒店(北京天安门店)')&lt;=hotel_names_set</pre>	<b>Query:</b> My brother and I are planning to travel from Shanghai to Chongqing for 4 days. <b>Apart from the round-trip high-speed train/flight, we aim to spend no more than 3400 yuan in Chongqing</b> . <b>DSL Translation (GPT-4o):</b> <pre>total_cost=0 for activity in allactivities(plan):     total_cost+=activity_cost(activity)+innercity_transport_cost(activity_transports(activity)) result=(total_cost&lt;=3400)</pre>

Figure 4: Challenges in the Neuro-Symbolic Planning.

<pre># Dining expenses &lt;= 1000 CNY. dining_cost = 0 for act_i in allactivities(plan):     typ = activity_type(act_i)     if typ=="breakfast" or typ=="lunch"     or typ=="dinner": dining_cost =     dining_cost + activity_cost(act_i) return dining_cost &lt;= 1000</pre>	<pre># Arriving in Shanghai should be before 6 PM on the second day. return_time = 0 for act_i in day_activities(plan, 2):     typ = activity_type(act_i)     dest = transport_destination(act_i)     if (typ=="train" or typ=="airplane")     and des=="Shanghai": return_time ==     activity_endtime(act_i) return return_time &lt; "18:00"</pre>	<pre># The number of attractions visited count = 0 for act_i in all_activities(plan):     if     activity_type(act_i)=="attraction":     count = count + 1 return count # Compare the return during evaluation of preference ranking</pre>
(a) Dining expenses.	(b) Arrived Time.	(c) Count of attraction visited.

Figure 5: Examples of DSL expressions for logical constraints and preference ranking.

346 **Core Concept Functions** We have defined 35 concept functions. Their definition and implementa-  
347 tion is in Table 9, 10, 11 and 12. Below are common use cases:

348 Example: Budget Constraint User Query: "Total expenses must not exceed 5000 CNY."

```
349
350 total_cost = 0
351 for act in all_activities(plan):
352     total_cost += activity_cost(act)
353     total_cost += innercity_transport_cost(activity_transports(act))
354 return total_cost <= 5000
```



Tool	API	Docs
Attractions	attractions_keys(city)	Return a list of (key, type) pairs of the attractions data.
	attractions_select(city, key, func)	Return a DataFrame with data filtered by the specified key with the specified function.
	attractions_id_is_open(city, id, time)	Return whether the attraction with the specified ID is open at the specified time.
	attractions_nearby(city, point, topk, dist)	Return the top K attractions within the specified distance of the location.
	attractions_types	Return a list of unique attraction types.
Accommodations	accommodations_keys(city)	Return a list of (key, type) pairs of the accommodations data.
	accommodations_select(city, key, func)	Return a DataFrame with data filtered by the specified key with the specified function.
	accommodations_nearby(city, point, topk, dist)	Return the top K accommodations within the specified distance of the location.
Restaurants	restaurants_keys(city)	Return a list of (key, type) pairs of the restaurants data.
	restaurants_select(city, key, func)	Return a DataFrame with data filtered by the specified key with the specified function.
	restaurants_id_is_open(city, id, time)	Return whether the restaurant with the specified ID is open at the specified time.
	restaurants_nearby(city, point, topk, dist)	Return the top K restaurants within the specified distance of the location.
	restaurants_with_recommended_food(city, food)	Return all restaurants with the specified food in their recommended dishes.
	restaurants_cuisine(city)	Return a list of unique restaurant cuisines.
Transportation	goto(city, start, end, start_time, transport_type)	Return a list of transportation options between two locations with the specified departure time and transportation mode.
IntercityTransport	intercity_transport_select(start_city, end_city, intercity_type, earliest_leave_time)	Return the intercity transportation information between two cities.
Others	notedown(description, content)	Write the specified content to the notebook
	plan(query)	Generates a plan based on the notebook content and query and report the plan is done.
	next_page()	Get the next page of the latest Result history if it exists. Because of the length limited, all returned DataFrame information is split into 10 rows per page.

Table 4: APIs

Category	Environment Constraints	Semantics
Cross-city Transportation	Intercity transportation events must occur.	The first event and last event must be cross-city transports.
	Available Trains or Airplanes across cities.	The provided TrainID/FlightID, origin and destination should be valid in the travel sandbox.
	Correct information of price, duration.	The price and duration information should match the travel sandbox.
	Detailed cost on inter-city transportation	Provide number of tickets and cost of each inter-city activity. $cost = price \times tickets$
Inner-city Transportation	Available Metro, Taxi or Walking between different positions.	The provided routes should be valid in the travel sandbox.
	Correct information of price, distance, and duration.	These details should match the travel sandbox.
	Detailed cost on inner-city transportation	Provide number of tickets/cars and cost. Taxi: 4 people per car. $cost = price \times tickets$ , $cost = price \times cars$
Attractions	Available attractions in the target city	The provided attractions should be valid in the travel sandbox.
	Visiting during opening hours.	Activities must respect the attraction's opening time.
	Correct price information.	Must match the sandbox.
	Detailed cost of attraction activity.	Provide ticket number and total cost. $cost = price \times tickets$
Restaurants	No repeated attractions.	Attractions should not repeat across the trip.
	Available restaurants in the target city	Must be valid in the travel sandbox.
	Visiting during opening hours.	Same as above.
	Correct price information.	Must match the sandbox.
Restaurants	Detailed cost of restaurant activity.	$cost = price \times tickets$
	No repeated restaurants.	Same restaurant should not be visited more than once.
	Meals served in proper time slots.	Breakfast: 06:00–09:00; Lunch: 11:00–14:00; Dinner: 17:00–20:00.
Accommodation	Available accommodations in target city.	Must be valid in the travel sandbox.
	Correct price and room type.	Must match the sandbox.
	Detailed accommodation cost.	$cost = price \times rooms$
Time	Required for trips over one day.	A hotel is necessary for multi-day trips.
	Activity duration details.	Must include start and end time; end time must be after start.
	Activities in chronological order.	Events listed in order, respecting preceding transport arrivals.
Space	Transport info for changing positions.	If positions differ, the transport route must be included.

Table 5: Environment Constraints and Semantics in ChinaTravel Environment

356 The function `all_activities(plan)` iterates through all activities in the itinerary. The function `activity_cost` retrieves the cost of each activity. The function `innercity_transport_cost` sums transportation  
357 expenses. Based on Python syntax, combining these concept functions can calculate the cost of the  
358 entire plan, thereby determining whether the budget constraints are met.  
359

360 **Debugging Tips** (1) Syntax Validation: Use the python compiler to check for syntax errors (e.g.,  
361 missing colons, undefined variables). (2) Unit Testing: Test individual concept functions (e.g.,  
362 `poi_distance`) with mock itineraries. (3) Iterative Refinement: For ambiguous requirements (e.g.,

<i>Logical Constraint</i>	
Transportation	The required type of transportation.
Attraction	The required type or specified attractions.
Restruants	The required type or specified restruants.
Accommodation	The number of rooms and the room type must meet the requirements. The required features or specified hotels.
Budget	The total cost is within required budget.
<i>Unseen Logical Constraints in Human data</i>	
POIs	The negation/conjunction/disjunction of given POIs
Time	The duration of specific activities is within the limitation
Budget	The cost of specific activities is within the required budget

Table 6: Descriptions of **Logical Constraints** for two benchmarks. Constraints in black are common in both TravelPlanner and ChinaTravel. Metrics in brown are the metrics only in our benchmark.

<i>Preference Requirements</i>	
Daily attractions ↑	Visit as many attractions as possible
Transport time ↓	Minimize the travel time between POIs
Transport time to the restaurants ↓	Minimize the travel time to restaurants
Food cost ratio ↑	Maximize the proportion of dining expenses
Hotel cost ↓	Minimize accommodation costs
Distance to POI ↓	Visit places as close to {POI} as possible

Table 7: Descriptions of **Preference Requirements** in ChinaTravel benchmark.













363 "local cuisine"), map natural language to precise DSL concepts from sandbox information (e.g.,  
364 restaurant\_type(act, city) == "Beijing Cuisine").

365 **Integration with Neuro-Symbolic Agents.** (1) NL2DSL Translation: Convert user queries into  
366 DSL using LLMs (e.g., "Try local food" → restaurant\_type(POI, city) == "Beijing Cuisine" when the  
367 destination city is Beijing). (2) Symbolic Validation: Execute DSL code to verify plans against logical  
368 constraints. (3) Search Optimization: Use DSL-defined preferences (e.g., minimize(transport.time))  
369 to rank candidate itineraries.

### 370 D.3 Query Synthesis

371 We designed common travel information (origin, destination, days, number of people) and logical  
372 constraints based on the nature of travel tasks. To facilitate scalable queries for ChinaTravel, we  
373 randomly constructed query skeletons from the aforementioned information and used advanced LLMs  
374 to generate natural language queries from these skeletons. In practice, we provide the LLMs with  
375 more intuitive hard logic constraints to ensure the LLMs do not make mistakes and use a Python  
376 script to convert it after generating the query. The automatically generated data is categorized into

Table 8: Results of different LLMs and planning strategies on the ChinaTravel *medium* subset.

		DR		EPR		LPR		C-LPR	FPR		DR		EPR		LPR		C-LPR	FPR
				Mic.	Mac.	Mic.	Mac.						Mic.	Mac.	Mic.	Mac.		
Act		72.7	52.3	0	63.5	15.3	0	0	0	NSP		71.3	71.9	69.3	69.4	50.0	69.3	46.7
		97.4	70.5	0	89.3	55.3	0	0	0			68.0	68.0	68.0	64.1	46.6	64.1	46.7
ReAct (zero-shot)		41.3	35.2	0	37.6	4.0	0	0	0	NSP		53.3	45.9	16.0	49.2	33.3	14.8	8.50
		92.0	54.8	0	78.6	22.7	0	0	0			68.6	65.4	54.0	66.2	61.3	52.5	54.0
ReAct (one-shot)		82.7	77.1	3.33	82.6	48.7	2.95	1.33	1.33	oracle		60.8	59.4	54.9	60.3	58.2	60.3	56.9
		94.7	69.2	0.67	91.8	64.0	0.53	0	0			53.3	51.3	36.6	51.9	43.3	34.8	34.6

Function Name	Meaning	Implementation
day_count	total days in the plan	<pre>def day_count(plan):     return len(plan["itinerary"])</pre>
people_count	number of people in the trip	<pre>def people_count(plan):     return plan["people_number"]</pre>
start_city	start city of the plan	<pre>def start_city(plan):     return plan["start_city"]</pre>
target_city	target city of the plan	<pre>def target_city(plan):     return plan["target_city"]</pre>
allactivities	all the activities in the plan	<pre>def allactivities(plan):     activity_list = []     for day_activity in plan["itinerary"]:         for act in day_activity["activities"]:             activity_list.append(act)     return activity_list</pre>
allactivities_count	the number of activities in the plan	<pre>def allactivities_count(plan):     count = 0     for day_activity in plan["itinerary"]:         count += \             len(day_activity["activities"])     return count</pre>
dayactivities	all the activities in the specific day [1, 2, 3, ...]	<pre>def dayactivities(plan, day):     activity_list = []     for act in plan["itinerary"]\         [day - 1]["activities"]:         activity_list.append(act)     return activity_list</pre>
activity_cost	the cost of specific activity without transport cost	<pre>def activity_cost(activity):     return activity.get("cost", 0)</pre>
activity_position	the position name of specific activity	<pre>def activity_position(activity):     return activity.get("position", "")</pre>
activity_price	the price of specific activity	<pre>def activity_price(activity):     return activity.get("price", 0)</pre>
activity_type	the type of specific activity	<pre>def activity_type(activity):     return activity.get("type", "")</pre>
activity_tickets	the number of tickets needed for specific activity	<pre>def activity_tickets(activity):     return activity.get("tickets", 0)</pre>
activity_transports	the transport information of specific activity	<pre>def activity_transports(activity):     return activity.get("transports", [])</pre>
activity_start_time	the start time of specific activity	<pre>def activity_start_time(activity):     return activity.get("start_time")</pre>
activity_end_time	the end time of specific activity	<pre>def activity_end_time(activity):     return activity.get("end_time")</pre>

Table 9: Concept Function

Function Name	Meaning	Implementation
activity_time	the duration of specific activity	<pre> def activity_time(activity):     start_time = activity.get("start_time")     end_time = activity.get("end_time")     if start_time and end_time:         st_h, st_m = \             map(int, start_time.split(":"))         ed_h, ed_m = \             map(int, end_time.split(":"))         return \             (ed_m - st_m) + (ed_h - st_h) * 60     return -1 </pre>
poi_recom- mend_time	the recommend time of specific poi(attraction) in the city	<pre> def poi_recommend_time(city, poi):     select = Attractions().select     attrction_info = \         select(city, key="name",             func=lambda x: x == poi).iloc[0]     recommend_time = \         (attrction_info["recommendmintime"]) \         * 60     return recommend_time </pre>
poi_distance	the distance between two POIs in the city	<pre> def poi_distance(city, poi1, poi2):     start_time="00:00"     transport_type="walk"     goto = Transportation().goto     return goto(city, poi1, poi2, start_time,         transport_type)[0]["distance"] </pre>
innercity_- transport_cost	the total cost of specific innercity transport	<pre> def innercity_transport_cost(transports, mode):     cost = 0     for transport in transports:         if mode is None or \             transport.get("type") == mode:             cost += transport.get("cost", 0)     return cost </pre>
innercity_- transport_price	the price of innercity transport	<pre> def innercity_transport_price(transports):     price = 0     for transport in transports:         price += transport["price"]     return price </pre>
innercity_- transport_- distance	the distance of innercity trans- port	<pre> def innercity_transport_distance\ (transports, mode=None):     distance = 0     for transport in transports:         if mode is None or \             transport.get("type") == mode:             distance += \                 transport.get("distance", 0)     return distance </pre>
innercity_- transport_- time	the duration of innercity trans- port	<pre> def innercity_transport_time(transports):     def calc_time_delta(end_time, start_time):         hour1, minu1 = \             int(end_time.split(":")[0]), \             int(end_time.split(":")[1])         hour2, minu2 = \             int(start_time.split(":")[0]), \             int(start_time.split(":")[1])         return (hour1 - hour2) * 60\             + (minu1 - minu2) </pre>

Table 10: Concept Function

Function Name	Meaning	Implementation
metro_tickets	the number of metro tickets if the type of transport is metro	<pre>def metro_tickets(transport):     return transports[1]["tickets"]</pre>
taxi_cars	the number of taxi cars if the type of transport is taxi	<pre>def taxi_cars(transport):     return transports[0]["cars"]</pre>
room_count	the number of rooms of accommodation	<pre>def room_count(activity):     return activity.get("rooms", 0)</pre>
room_count	the number of rooms of accommodation	<pre>def room_count(activity):     return activity.get("rooms", 0)</pre>
room_type	the type of room of accommodation	<pre>def room_type(activity):     return activity.get("room_type", 0)</pre>
restaurant_type	the type of restaurant's cuisine in the target city	<pre>def restaurant_type(activity, target_city):     restaurants = Restaurants()     select_food_type = \         restaurants.select(             target_city, key="name",             func=lambda x: x == activity["position"]         )["cuisine"]     if not select_food_type.empty:         return select_food_type.iloc[0]     return ""</pre>
attraction_type	the type of attraction in the target city	<pre>def attraction_type(activity, target_city):     attractions = Attractions()     select_attr_type = \         attractions.select(             target_city, key="name",             func=lambda x: x == activity["position"]         )["type"]     if not select_attr_type.empty:         return select_attr_type.iloc[0]     return ""</pre>
accommodation_type	the feature of accommodation in the target city	<pre>def accommodation_type(activity, target_city):     accommodations = Accommodations()     select_hotel_type = \         accommodations.select(             target_city, key="name",             func=lambda x: x == activity["position"]         )["featurehoteltype"]     if not select_hotel_type.empty:         return select_hotel_type.iloc[0]     return ""</pre>
innercity_transport_type	the type of innercity transport	<pre>def innercity_transport_type(transport):     if len(transport) == 3:         return transports[1]["mode"]     elif len(transport) == 1:         return transports[0]["mode"]     return ""</pre>
intercity_transport_type	the type of intercity transport	<pre>def intercity_transport_type(activity):     return activity.get("type", "")</pre>

Table 11: Concept Function



<p>Query in Chinese (from easy subset):当前位置成都。我和朋友两个人想去南京玩 2 天，住一间双床房，酒店可以打牌，请给我一个旅行规划。</p> <p>Current location: Chengdu. My friend and I want to go to Nanjing for 2 days. We need a twin room in a hotel where we can play cards. Please provide a travel plan for us.</p>
<pre> accommodation_type_set=set() for activity in allactivities(plan):     if activity_type(activity) == 'accommodation': accommodation_type_set.add(accommodation_type(activity, target_city(plan))) result=({'棋牌室'}&lt;=accommodation_type_set) </pre>
<p>Query in Chinese (from medium subset): 当前位置成都。我一个人想去重庆玩 2 天，预算 3000 人民币，坐火车往返，想吃火锅，想去洪崖洞。</p> <p>Current location: Chengdu. I want to travel alone to Chongqing for 2 days with a budget of 3000 RMB. I plan to take the train, want to eat hotpot, and visit Hongya Cave.</p>
<pre> total_cost=0 for activity in allactivities(plan):     total_cost+=activity_cost(activity)     total_cost += innercity_transport_cost(activity_transports(activity)) result=(total_cost&lt;=3000) intercity_transport_set=set() for activity in allactivities(plan):     if activity_type(activity) in ['train', 'airplane']: intercity_transport_set.add(intercity_transport_type(activity)) result=({'train'}==intercity_transport_set)" restaurant_type_set=set() for activity in allactivities(plan):     if activity_type(activity) in ['breakfast', 'lunch', 'dinner']: restaurant_type_set.add(restaurant_type(activity, target_city(plan))) result=({'火锅'}&lt;=restaurant_type_set) attraction_name_set=set()\nfor activity in allactivities(plan):     if activity_type(activity)=='attraction': attraction_name_set.add(activity_position(activity)) result=({'洪崖洞'}&lt;=attraction_name_set) </pre>
<p>Query in Chinese (from human subset): [当前位置南京,目标位置武汉,旅行人数 2,旅行天数 3] 我们 2 人想去武汉玩 3 天，主要想体验武汉的一些有些历史的区域，同时还想尝一尝本地人常去吃的特色美食，怎么规划行程。</p> <p>English translation: [Current location: Nanjing, Destination: Wuhan, Number of travelers: 2, Travel days: 3] The two of us want to visit Wuhan for 3 days. We mainly want to experience some of the historical areas in Wuhan and also try the local specialty foods that residents often eat. How should we plan our itinerary?</p>
<pre> attraction_type_set=set() for activity in allactivities(plan):     if activity_type(activity)=='attraction': attraction_type_set.add(attraction_type(activity, target_city(plan))) result=({'历史古迹'}&lt;=attraction_type_set)" restaurant_type_set=set()\nfor activity in allactivities(plan):     if activity_type(activity) in ['breakfast', 'lunch', 'dinner']: restaurant_type_set.add(restaurant_type(activity, target_city(plan))) result=({'湖北菜'}&lt;=restaurant_type_set)" </pre>
<p>Query in Chinese (from human subset): [当前位置南京,目标位置杭州,旅行人数 2,旅行天数 3] 我们打算去杭州看西湖，预算 2000，给我一个旅游安排。</p> <p>[Current location: Nanjing, Destination: Hangzhou, Number of travelers: 2, Number of travel days: 3] We plan to visit West Lake in Hangzhou with a budget of 2000. Please provide me with a travel itinerary.</p>
<pre> attraction_name_set=set() for activity in allactivities(plan):     if activity_type(activity)=='attraction': attraction_name_set.add(activity_position(activity)) result=({'西湖风景名胜区'}&lt;=attraction_name_set) total_cost=0 for activity in allactivities(plan):     total_cost+=activity_cost(activity)     total_cost += innercity_transport_cost(activity_transports(activity)) result=(total_cost&lt;=2000)" </pre>

Figure 6: Examples of travel requirements and their DSL expressions.

Function Name	Meaning	Implementation
innercity_-transport_-start_time	the start time of innercity transport	<pre>def innercity_transport_start_time(transports):     return transports[0]["start_time"]</pre>
innercity_-transport_-end_time	the end time of innercity transport	<pre>def intercity_transport_end_time(transports):     return transports[-1]["end_time"]</pre>
intercity_-transport_-origin	the origin city of intercity transport	<pre>def intercity_transport_origin(activity):     if "start" in activity:         for city in city_list:             if city in activity["start"]:                 return city     return ""</pre>
intercity_-transport_-destination	the destination city of intercity transport	<pre>def intercity_transport_destination(activity):     if "end" in activity:         for city in city_list:             if city in activity["end"]:                 return city     return ""</pre>

Table 12: Concept Function

377 two difficulty levels: In the *Easy* level, user inputs encompass a single logical requirement, sourced  
 378 from categories such as transportation, restaurants, attractions, and accommodations. In the *Medium*  
 379 level, user inputs involve 2 to 5 logical requirements, introducing more complex constraints. During  
 380 the generation, we encourage the LLMs to provide varied and human-like expressions, necessitating a  
 381 deeper understanding and processing to accurately interpret and fulfill the user’s needs. For instance,  
 382 the logical requirement “taste Beijing cuisine” could correspond to the natural language query: “Try  
 383 local food in Beijing.” We utilize prompt engineering to guide LLMs in refining natural language  
 384 expressions to facilitate automated generation. One of the prompts is shown in Figure 7. Several  
 385 examples of generated data is in Figure 8. As a result, we obtain the synthetic queries across diverse  
 386 travel requirements, including 28 restaurant types, 23 attraction types, 29 hotel features, and more  
 387 than 130 specific POIs.

## D.4 Data Diversity and Bias Mitigation

This subsection provides a detailed analysis of ChinaTravel’s hybrid query design, addressing concerns about synthetic data limitations and real-world representativeness.

ChinaTravel integrates both synthetic and human-authored queries to balance scalability and realism. When synthesizing data, we randomly constructed constraints based on the types and specific visit requirements of POIs such as restaurants, accommodations, transports, and attractions, thereby ensuring the diversity of the dataset. The human query subset comprises 154 samples collected through structured questionnaires, which introduce complex real-world constraints such as time-bound returns (e.g., explicit requirements like “return before 7 PM”) and activity-specific budget allocations. These queries also incorporate colloquial expressions that reflect native Chinese travel preferences, exemplified by phrases like local specialty foods frequented by residents. The synthetic queries are generated through LLM-based paraphrasing techniques and systematically categorized into two tiers: Easy-tier queries contain single logical constraints (e.g., specific cuisine requirements), while Medium-tier queries combine 3–5 interdependent constraints (e.g., compound conditions like “budget  $\leq$  3000 CNY + train transport + hotpot dining”).

To mitigate synthetic data bias and enhance diversity, three primary strategies were implemented. First, constraint combinations were deliberately diversified across temporal, spatial, and cost dimensions, as detailed in Table 6. Second, a human validation layer filters out unrealistic LLM-generated queries, such as physically implausible itineraries like “visiting 10 attractions within one day.” Third, the DSL framework enables compositional generalization of requirements, supporting open-ended constraint combinations through its formal syntax shown in Table 1.

The current human query subset remains limited by annotation costs, as discussed in the limitation section. In future work, we will advance data collection by integrating LLMs with real user queries to automate and diversify the generation of human-like queries. Additionally, all human queries and automated synthesis tools will be publicly released to support community-driven benchmark extensions.

## D.5 Data with Preference

We introduce six common preferences from user surveys to construct the preference sub-datasets. Table 7 provides a summary of these preferences.

The corresponding DSL could be formulated as follows.

```
# The number of attractions visited
count = 0
for act_i in all_activities(plan):
    if activity_type(act_i)=="attraction": count = count + 1
return count
```

```
# The average travel time between POIs
time_cost = 0
transport_count = 0
for activity in allactivities(plan):
    transports = activity_transports(activity)
    transport_count += 1
    time_cost += innercity_transport_time(transports)
average_time_cost = time_cost / transport_count if transport_count > 0
else -1
return average_time_cost
```

```
# The average travel time to restaurants
restaurant_count = 0
time_cost = 0
for activity in allactivities(plan):
    if activity_type(activity) in ['breakfast', 'lunch', 'dinner']:
        restaurant_count += 1
        time_cost += innercity_transport_time(activity_transports(
            activity))
```

```

446 if restaurant_count == 0:
447     average_time_cost = -1
448 else:
449     average_time_cost = time_cost / restaurant_count
450 return average_time_cost
451
452
453 # The ratio of food cost
454 food_cost = 0
455 total_cost = 0
456 for activity in allactivities(plan):
457     total_cost += activity_cost(activity)
458     total_cost += innercity_transport_cost(activity_transports(activity))
459     if activity_type(activity) in ['breakfast', 'lunch', 'dinner']:
460         food_cost += activity_cost(activity)
461 food_cost_ratio = food_cost / total_cost if total_cost > 0 else -1
462 return food_cost_ratio
463
464
465 # The cost of accommodations
466 accommodation_cost = 0
467 for activity in allactivities(plan):
468     if activity_type(activity) == 'accommodation':
469         accommodation_cost += activity_cost(activity)
470 return accommodation_cost
471
472
473 # The average distance to poi (e.g. xxx)
474 target_poi = 'xxx'
475 poi_list = list()
476 total_distance = 0
477 poi_count=0
478 city = target_city(plan)
479 for activity in allactivities(plan):
480     if activity_type(activity) in ['breakfast', 'lunch', 'dinner', 'accommodation', 'attraction']:
481         poi_list.append(activity_position(activity))
482 for poi in poi_list:
483     total_distance += poi_distance(city, target_poi, poi)
484     poi_count += 1
485 average_dist_cost = total_distance / poi_count if poi_count > 0 else -1
486 return average_dist_cost
487
488
489

```

## 491 D.6 Benchmark Difficulty and Applicability

492 While the Human subset presents significant challenges, the baseline NeSy solution has achieved  
493 60.6% and 46.7% FPR on Easy and Medium subsets, respectively, providing developers with action-  
494 able validation points for initial testing and refinement. Additionally, the Human subset’s extreme  
495 difficulty arises from open language reasoning and unseen concept composition, key challenges  
496 absent in prior benchmarks but unavoidable in practice. By explicitly formalizing these challenges,  
497 ChinaTravel has provided a roadmap for advancing agents toward real-world robustness. Despite  
498 current LLMs’ limitations in handling unseen combinations, their success in code generation suggests  
499 that post-training with DSL may enhance their understanding of diverse travel needs, moving toward  
500 real-world applications.

## 501 E Discussion with TravelPlanner

502 TravelPlanner’s logical constraints contain the total cost, 15 cuisines, 5 house rules, 3 room types,  
503 and 3 intercity transports. ChinaTravel’s logical constraints contain the total cost, 42 cuisines, 15  
504 attraction types, 78 hotel features, 2 room types, 2 intercity-transports types, 3 inner-city-transports  
505 types, and specific POI names (attractions, restaurants, hotels). Crucially, our benchmark introduces

compositional constraints derived from human queries (e.g., “return before 7 PM”, “cost of intercity transports”), reflecting real-world complexity. The key advancement lies in addressing open-language reasoning and unseen concept composition, which represent significant challenges beyond TravelPlanner’s scope. Our Domain-Specific Language (DSL) enables automated validation of these combinatorial requirements, bridging the gap between synthetic and real-world needs.

We also provide some example queries and corresponding examples from the TravelPlanner at each level in Figure 18, 19, and 20.

As shown in Figure 18, in the “easy level”, TravelPlanner only includes constraints on cost. In contrast, ChinaTravel demonstrates significant advantages over TravelPlanner, particularly in terms of personalized support for specific Points of Interest (POIs) and more realistic transportation and time management. These advantages are crucial for developing and evaluating language agents that can handle real-world travel planning scenarios effectively. ChinaTravel allows users to directly specify POI names, such as “Nanjing DaPaXiang” or “HuQiu Mountain Scenic Area,” requiring the agent to precisely match the entity information from the travel sandbox.

As shown in Figure 19, in the medium set, TravelPlanner includes queries with two types of constraints: cost and cuisine, or cost and accommodation. In contrast, ChinaTravel includes queries with 2 to 5 types of constraints, reflecting more complex and diverse multi-constraint requirements. This difference highlights the ability of ChinaTravel to handle more realistic and varied travel planning scenarios.

As shown in Figure 20, TravelPlanner includes queries with multiple constraints, such as cost, accommodation type, and cuisine preferences. However, ChinaTravel goes a step further by including queries with unseen logical constraints and more colloquial expressions. These unseen logical constraints and colloquial expressions are essential for travel planning agents to handle real-world users effectively. They reflect the complexity and diversity of real-world travel planning scenarios, where users may have diverse requirements that need to be understood and addressed. By incorporating these elements, ChinaTravel bridges the gap between current academic research benchmarks and real-world application demands, making it a more comprehensive and realistic benchmark for evaluating the capabilities of travel planning agents.

## F NeSy Planning

Since the Z3 solver from [10] would restructure the tool API to return travel information expressed in specific Z3 variables, which may not be feasible given that APIs in the real world are typically black boxes that agents can only call. Following their two-stage solution, we first extract logical constraints from natural language. Based on these constraints, we implement a step-by-step plan generation process using depth-first search, mimicking how humans plan to travel by arranging activities one by one. As shown in Fig. 9, we first translate the natural languages to logical constraints through prompting. generate the next activity type based on the current plan, and then recursively generate the next activity until the goal is reached. The generated plan is then used to solve the problem. In the second step, we define the rule-based activity selection and score function. For example, if the current time is in the [10:30, 12:30] and there is no scheduled lunch in the current plan, then the agent should find a restaurant to have lunch at this time. If the current time is after 22:00 and there are no open-time attractions nearby, the agent should choose to return to the hotel. For the score function, we select the restaurants that satisfy the required cuisine and sort the candidates by the price if there a budget constraints in the constraints  $C$ . These ranking functions will help us to find a feasible solution as soon as possible. In ChinaTravel, the duration arrangement of activities is continuous and difficult to

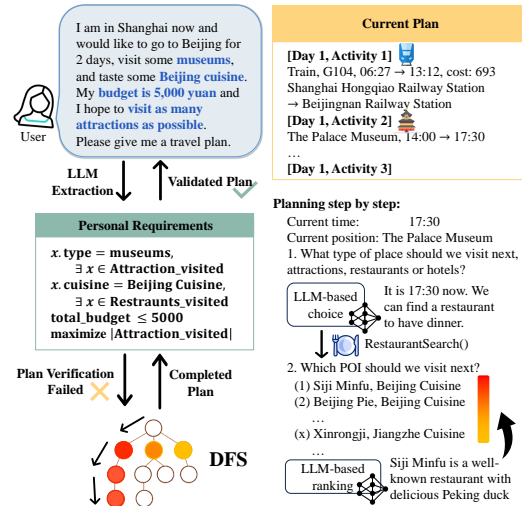


Figure 9: NeSy Planning with search-based solver.

---

**Algorithm 1** Depth-First Greedy Search

---

**Require:** Constraints  $C$ , current plan  $p$ ,  
if the least activity is an intercity-transport from destination to origin **then**  
    **return** ConstraintValidation( $p$ ,  $C$ ),  $p$        $\triangleright$  The plan  $p$  is finished, return the validation result.  
**end if**  
type = GetNextActivityType( $p$ )       $\triangleright$  Select the next type of activities, e.g. lunch, attraction.  
candidates = ToolUse(type)       $\triangleright$  Collect the corresponding information for the activity type  
scores = LLMScore(candidates,  $p$ ,  $C$ )       $\triangleright$  Score candidates through constraints  $C$ .  
**for** activity in candidates **do**  
     $p$ .push(activity)       $\triangleright$  Perform a greedy search with priority ranking.  
    flag,  $p$  = Depth-FirstGreedySearch( $C$ ,  $p$ )  
    **if** flag **then**  
        **return** True,  $p$        $\triangleright$  Return the solution  $p$  if the validation is passed.  
    **end if**  
     $p$ .pop(activity)  
**end for**  
**return** False,  $p$        $\triangleright$  Fail to find a solution with the given conditions.

---

560 enumerate and search. We pre-define a meal or a visit to an attraction as 90 minutes, and when there  
561 are less than 90 minutes until closing time, the event continues until the closing time. Given these  
562 designs, we adapt the neural-symbolic solution into a multi-POI planning problem and evaluate it in  
563 the ChinaTravel benchmark.

564 Given that some queries are particularly challenging due to the limited number of feasible plans,  
565 we set the maximum runtime for the symbolic sketch from interactive search to 5 minutes per  
566 query, excluding the LLM inference time, to ensure a fair comparison across different models. If  
567 a plan satisfying the generated DSL validation is found within the time limit, it is returned directly.  
568 Otherwise, the program halts when the time limit is reached, and the plan that satisfies environmental  
569 constraints while achieving the highest number of validation code successes among all intermediate  
570 results is returned. In cases where no environment-compliant plan is identified, the partially completed  
571 plan generated up to that point is returned.

572 In the Figure 21, 22 and 23, we provide the prompts of the LLM POI-ranking phases.

## 573 G Evaluation Metric in Competition

574 The Delivery Rate (DR), Environmental Pass Rate (EPR), Logical Pass Rate (LPR), and Final Pass  
575 Rate (FPR) have been detailed in TravelPlanner [24]. To make the paper more self-contained, we  
576 also provide the corresponding definition here.

577 **Delivery Rate:** This metric assesses whether agents can successfully deliver a formatted plan. For  
578 the Nesy planning, if a solution that satisfies the logical constraints has not been found by the time  
579 out, the search is terminated, and the current solution that satisfies the environmental constraints  
580 is returned. If no solution that satisfies the environmental constraints is obtained, an empty plan  
581 is returned. Therefore, unlike the pure LLM method, which primarily assesses the Delivery Rate  
582 based on whether the output meets the formatting requirements, the nesy planning method, which  
583 uses depth-first-search to arrange POIs one by one, shows differences in the Delivery Rate. These  
584 differences mainly reflect the proportion of effective solutions obtained within a limited time based  
585 on the LLM’s POI recommendation. This proportion demonstrates the degree to which the LLM  
586 prioritizes POI arrangements from a natural language perspective and meets formalized logical  
587 requirements. The more accurately the LLM can arrange POIs that are beneficial for long-horizon  
588 planning, the more likely it is to obtain effective solutions and improve the Delivery Rate.

589 **Environmental Pass Rate** Comprising the environmental dimensions (as detailed in Tab. 5), this  
590 metric evaluates whether a language agent can accurately incorporate sandbox information into their  
591 generated plans.



$$EPR - micro = \frac{\sum_{p \in P} \sum_{c \in Env} \mathbb{1}_{passed(c,p)}}{|P| * |Env|}$$

$$EPR - macro = \frac{\sum_{p \in P} \prod_{c \in Env} \mathbb{1}_{passed(c,p)}}{|P|}$$

**Logical Pass Rate** Comprising the logical dimensions (as detailed in Tab. 6), this metric evaluates whether a language agent can accurately meet the personalized requirements of the queries.

$$LPR - micro = \frac{\sum_{p \in P} \sum_{c \in C_p} \mathbb{1}_{passed(C_p,p)}}{\sum_{p \in P} |C_p|}$$

$$LPR - macro = \frac{\sum_{p \in P} \prod_{c \in C_p} \mathbb{1}_{passed(C_p,p)}}{|P|}$$

**Final Pass Rate** This metric represents the proportion of feasible plans that meet all aforementioned constraints among all tested plans. It serves as an indicator of agents' proficiency in producing plans that meet a practical standard.

$$FPR = \frac{\sum_{p \in P} \mathbb{1}_{passed(Env,p)} \cdot \mathbb{1}_{passed(C_p,p)}}{|P|}$$

**Preference Ranking** To systematically evaluate the satisfaction of soft user preferences in travel planning, we introduce a Preference Ranking metric that quantifies the alignment of generated itineraries with diverse user requirements. Each preference (e.g., "maximize attraction visits" or "minimize transportation time") is formalized into a Domain-Specific Language (DSL)-based concept, enabling automated numerical extraction from plans. For instance, the preference for "visiting more attractions" is translated into a DSL function that counts the total attraction-type activities in a plan, while "minimizing dining costs" is operationalized via cumulative expense calculations for meal-related activities.

The Preference Ranking metric operates in two steps: 1) Quantification: Execute DSL code to compute concept-specific scores (e.g., attraction count, transport time) for each generated plan. 2) Ranking: Compare methods (e.g., BQ vs. PEQ vs. PDS) by ranking their concept values, prioritizing higher values for maximization goals ( $\uparrow$ ) and lower values for minimization goals ( $\downarrow$ ). 3) Aggregation: Calculate the average ranking on the given samples.

## H Detailed Empirical Analysis

### H.1 Main Results

Table 13: Cost per query across different methods.

Based on the results presented in Table 2 and 13, we have the following finding and analyses:

**Pure LLMs struggle in ChinaTravel.** The DR evaluates agents' capability to generate valid JSON travel plans (see Fig. 1). While high DR values indicate that state-of-the-art LLMs can produce structurally correct outputs, the near-zero EPR reveals their fundamental limitations in acquiring and strictly adhering to required constraints. The sole exception is the DeepSeek model, which achieves the 6% EPR and 5% FPR at easy level, likely due to its strong capability to follow Chinese requirements. ReAct (one-shot, GPT-4o) excels in Macro LPR but achieves no FPR, suggesting it circumvents constraints via shortcuts. Our proposed C-LPR metric offers a more reliable measure of logical constraints, serving as a supplement to FPR. As shown in Table 13, pure neural methods incur prohibitively high computational costs due to excessive token consumption. When powered by GPT-4o, each query incurs \$2.4 on average, yet these approaches fail to produce any constraint-satisfying plans.

Method	#Input	#Output	💎(\$)	🌱(\$)
Act	88K	2K	.007	.219
ReAct (0-shot)	206K	3K	.021	.638
ReAct (1-shot)	1058K	4K	.081	2.43
LLM-modulo	362K	11K	.025	1.12
NeSy Planning	467K	3K	.003	.306

**The Inadequacies of Existing NeSy Approaches.** The fundamental limitation of TTG arises from its computational complexity, where the constraint count scales as  $O(N^3T)$  with  $N$  POIs and  $T$  time windows. Even when subsampling to 22 POIs and discretizing time into 1-hour slots ( $T = 24$ ), the system generates approximately 600,000 constraints for 2-day itinerary. In our main experiments using the SCIP solver from the PuLP package, TTG was allocated a relaxed 15-minute search limitation. However, this configuration yielded only 18% valid solutions on easy-subet instances, with the FPR further reduced to 8% due to the solver’s pruning heuristics. Fig. 10a illustrates the solution time of TTG on 1-3 day itinerary. Within the time limit, solutions were found for merely 23% for two-day and 6% for three-day itineraries.

LLM-modulo introduces a oracle symbolic verifier and feedback the error to LLM to refine the plan. As illustrated in Fig. 10b, which depicts the error dynamics across 10 refinement iterations, GPT-4o maintains the lowest cumulative error count ( $\mu = 3.2 \pm 0.8$ ), followed by DeepSeek ( $\mu = 5.1 \pm 1.2$ ). However, their rectification capacity, quantified by successfully rectified errors per iteration rapidly decays to  $\leq 1$  after 3-5 rounds, indicating diminishing returns in error correction. Notably, smaller models (Qwen3-8B and Llama3-8B) achieve higher rectification, but this comes at the cost of introducing substantial emergent errors. The error reduction remains statistically insignificant across these models. This pattern suggests that while LLM-modulo enables basic constraint feedback from previous travel benchmarks [30, 24], its effectiveness diminishes for complex multi-day itineraries.

**Nesy Planning provides a promising solution.** Our NeSy Planning framework integrates symbolic programs to orchestrate travel planning and tool management while utilizing LLMs to extract language-based requirements and prioritize POIs. By separating understanding (flexible natural language handling), planning (DSL-guided backtrack) and grounding (precise execution), the framework enhances adaptability and ensures compliance with constraints during context-rich long-horizon planning. Across all data subsets, it outperforms previous TTG and LLM-modulo mthods, even without the help of oracle translation. Among the evaluated LLMs, DeepSeek-V3 achieves state-of-the-art performance across three subsets. With DeepSeek-V3 as the backend, it achieves FPRs of 52.6%, 37.0% and 23.3% on three subsets, highlighting the effectiveness of NeSy solutions for travel planning with complex constraints. Moreover, this superior performance demonstrates its enhanced capability for inter-constraint generalization in compositionally novel situations. Another potential explanation is that the model is developed by a Chinese company. As a result, it has been trained on a vast amount of Chinese-language data. This extensive exposure to Chinese text has enabled it to perform exceptionally well in our Chinese travel planning scenarios, giving it advantages over others.

**Challenges Persist for Nesy Planning.** The performance gap between standard and oracle modes underscores the importance of DSL translation in NeSy planning. Inadequate translations may result in plan searches failing to meet user requirements, while incorrect translations can misguide the search, making feasible solutions unattainable.

We conclude with three challenges and provide the corresponding cases in the Fig. 11. **(1) DSL Syntax Compliance:** As evidenced in Fig. 11a, while the reflexion process with syntactic validation effectively reduces surface-level errors, it inadvertently triggers constraint deletion behaviors across multiple LLMs. Specifically, Qwen3-8B, Llama3-8B, and Mistral-7B exhibit progressive reduction in extracted DSL constraints during iterative refinement. Notably, GPT-4o generates approximately two fewer constraints per iteration than DeepSeek-V3 on average. Although this conservative strategy enables rapid error convergence (achieving zero detected errors within limited iterations), it risks oversimplifying constraint specifications, critical dependencies may be prematurely discarded, ultimately yielding solutions that fail to satisfy complex requirements. This observed conservatism toward unseen constraints likely contributes to GPT-4o’s relative performance gap on the Human-154 and Human-1000 benchmarks compared to DeepSeek-V3. **(2) Open Contextual Reasoning:** In the Section 3 we have provided a quantitative analysis. In App. D.2, more examples are provied for this challenges. **(3) Unseen Concept Composition:** Real-world requirements are inherently diverse and complex, making expecting models to encounter all possible needs during development impractical. A more feasible way is to emulate human reasoning by generalizing existing knowledge to novel problems. Fig. 11b compares three LLMs on seen vs unseen DSL structures under POI-anonymized evaluation with syntax-level pattern matching. Results reveal critical gaps: 84% novel DSLs show only 12% alignment (9% overall), vs 93% accuracy on 16% known patterns. GPT-4o and Qwen3 also demonstrate this limitation, excelling on same concepts but failing on novel compositions.

In summary, ChinaTravel poses significant challenges for current agents. Neuro-symbolic agents outperform pure-LLM approaches in constraint satisfaction, showing strong potential for real-world travel planning. With realistic queries and a versatile DSL for constraint validation, we highlight the critical challenges while providing a foundation for advancing neuro-symbolic systems in practice.

## H.2 Ablation Study with Preference

The comparison of preferences should be conducted under the premise that both environmental and logical constraints are satisfied. Given the limited FPR achieved by existing methods, we perform a separate analysis of preference optimization here. Specifically, we sampled 50 queries from the easy subset that NeSy-DeepSeek-Oracle successfully passed as seed samples. Based on these, six subsets were created by introducing common preferences identified from user surveys. Three comparative scenarios were designed to explore the roles of LLMs and symbolic search in optimizing preferences during NeSy Planning: (1) BQ: Baseline solutions without preference consideration. (2) PEQ: LLM-enhanced recommendations with natural language preferences. (3) PDS: Hybrid symbolic search optimizing preference objectives under 5-min constraints. The results are provided in Fig. 12 (where  $\uparrow$  /  $\downarrow$  indicate maximization/minimization). We could find that: (1) PEQ outperforms BQ in 5/6 preference scenarios, confirming LLMs’ capacity to interpret natural language preferences during POI ranking. (2) PEQ underperforms on P2 (transport time minimization), likely from LLMs’ misinterpretation of complex spatiotemporal constraints. These results support the scalability of DSL in preference optimization but also highlights the pressing need for more efficient algorithms.

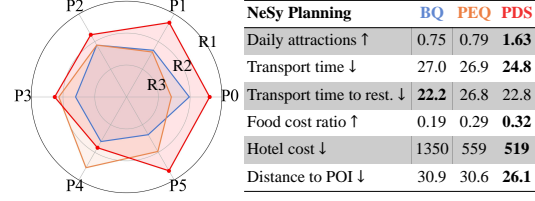


Figure 12: Ablation on preference ranking.

## I Additional Experimental Results

### I.1 Multi-Preference Comparison

For multi-preference scenarios (e.g., balancing "attraction visits  $\uparrow$ " and "transport time  $\downarrow$ "), we adopt an averaged aggregation approach, where rankings reflect the combined performance across all preferences. This framework ensures scalability and objectivity.

To rigorously evaluate the ability of language agents to balance multiple soft constraints, we constructed 15 test subsets by pairing six user preferences (P0–P5) into all possible combinations (e.g., "P0 + P1"). Each subset contains queries with two preference requirements. We compared two methods, Baseline Query (BQ) and Preference-Enhanced Query (PEQ), by quantifying their performance through our DSL-based Preference Ranking metric. For each subset, we extracted numerical scores for both preferences (Value-1 and Value-2), computed individual rankings (Rank-1, Rank-2), and derived an aggregated ranking (Agg. Rank.) to reflect overall performance. The results are provided in the Table 14.

From these results, we could find that: (1) **PEQ Outperforms BQ in Most Scenarios:** In 10/15 combinations, PEQ achieves superior aggregated rankings (Aggregated Ranking = 1.43 vs. BQ’s 1.56). Notably, PEQ demonstrates stable improvements on preferences P3 (e.g., maximizing dining quality $\uparrow$ ) and P4 (e.g., minimizing accommodation costs $\downarrow$ ). For instance: In "P0 $\uparrow$  + P4 $\downarrow$ ", PEQ reduces accommodation costs by 64% (Value-2: 441 vs. BQ’s 1221) while maintaining high attraction counts (Value-1: 0.97 vs. 0.79). For "P3 $\uparrow$  + P4 $\downarrow$ ", PEQ simultaneously improves dining quality (Value-1: 0.26 vs. BQ’s 0.18) and lowers costs (Value-2: 531 vs. 1229). This stability likely stems from the direct impact of POI selection on these preferences. LLMs in PEQ effectively prioritize low-cost hotels or high-quality restaurants through natural language hints (e.g., "reduce the cost on accommodations"), enabling explicit alignment with P3 and P4 requirements. (2) **Challenges in Balancing Multiple Preferences:** The results also reveal inherent difficulties in harmonizing conflicting preferences, particularly when optimizing one requirement necessitates sacrificing another. Notably, in the P0 $\uparrow$  + P1 $\downarrow$  scenario, PEQ underperforms BQ on both preferences, highlighting the inherent difficulty in resolving conflicting objectives. While PEQ marginally improves attraction counts (Value-1: 0.83 vs. BQ’s 0.79), it incurs a 5.7% increase in transport time (Value-2: 29.7 vs. BQ’s 28.0). This trade-off results in a worse aggregated ranking for PEQ (1.55 vs. BQ’s 1.44),

Preference Combination	Vaule-1		Vaule-2		Rank-1		Rank-2		Agg. Rank.	
	BQ	PEQ	BQ	PEQ	BQ	PEQ	BQ	PEQ	BQ	PEQ
P0 ↑, P1 ↓	0.79	<b>0.83</b>	<b>28.0</b>	29.7	<b>1.44</b>	1.55	<b>1.44</b>	1.55	<b>1.44</b>	1.55
P0 ↑, P2 ↓	0.82	<b>1.26</b>	<b>29.0</b>	31.9	1.56	<b>1.43</b>	<b>1.43</b>	1.56	1.5	1.5
P0 ↑, P3 ↑	0.81	<b>0.94</b>	0.18	<b>0.20</b>	<b>1.42</b>	1.57	1.59	<b>1.40</b>	1.51	<b>1.48</b>
P0 ↑, P4 ↓	0.79	<b>0.97</b>	1221	<b>441</b>	1.46	<b>1.53</b>	1.73	<b>1.26</b>	1.59	<b>1.40</b>
P0 ↑, P5 ↓	0.78	<b>0.91</b>	<b>33.6</b>	34.0	<b>1.37</b>	1.62	1.70	<b>1.29</b>	1.54	<b>1.45</b>
P1 ↓, P2 ↓	28.2	<b>27.8</b>	<b>26.6</b>	30.1	1.62	<b>1.37</b>	<b>1.48</b>	1.51	1.55	<b>1.44</b>
P1 ↓, P3 ↑	<b>28.2</b>	36.2	0.20	<b>0.27</b>	<b>1.31</b>	1.68	1.6	<b>1.4</b>	<b>1.45</b>	1.54
P1 ↓, P4 ↓	<b>30.3</b>	44.8	1440	<b>585</b>	<b>1.14</b>	1.85	1.77	<b>1.22</b>	<b>1.45</b>	1.54
P1 ↓, P5 ↓	<b>30.1</b>	38.3	30.7	<b>30.2</b>	<b>1.27</b>	1.72	1.69	<b>1.30</b>	<b>1.48</b>	1.51
P2 ↓, P3 ↑	24.7	<b>23.3</b>	0.27	0.27	<b>1.43</b>	1.56	1.60	<b>1.39</b>	1.52	<b>1.47</b>
P2 ↓, P4 ↓	24.1	<b>21.1</b>	1687	<b>719</b>	1.51	<b>1.48</b>	1.89	<b>1.10</b>	1.70	<b>1.29</b>
P2 ↓, P5 ↓	<b>28.0</b>	30.8	29.4	<b>26.0</b>	1.51	<b>1.48</b>	1.89	<b>1.10</b>	1.70	<b>1.29</b>
P3 ↑, P4 ↓	0.18	<b>0.26</b>	1229	<b>531</b>	1.64	<b>1.35</b>	1.69	<b>1.30</b>	1.66	<b>1.33</b>
P3 ↑, P5 ↓	0.22	0.22	33.3	<b>29.0</b>	1.51	<b>1.48</b>	1.84	<b>1.15</b>	1.68	<b>1.31</b>
P4 ↓, P5 ↓	1366	<b>767</b>	33.1	<b>31.6</b>	1.67	<b>1.32</b>	<b>1.45</b>	1.54	1.56	<b>1.43</b>
Aggregated Ranking									1.56	<b>1.43</b>

Table 14: Multi-Preference Comparison of BQ and PEQ.

indicating that the combined effect of conflicting preferences negates the benefits of natural language guidance. In 9/15 combinations, PEQ improves one preference at the expense of the other. For example: P1↓ + P4↓: PEQ reduces accommodation costs by 59% (Value-2: 585 vs. BQ’s 1440) but increases transport time by 48% (Value-1: 44.8 vs. 30.3). The inability to concurrently satisfy both preferences underscores the limitations of current LLM-driven prioritization in handling trade-offs.

Our experiments demonstrate that the neuro-symbolic agent (PEQ), enhanced by LLM-driven POI recommendation, outperforms baseline methods in multi-preference travel planning. By integrating natural language hints to guide POI selection, PEQ effectively translates user requirements into actionable itineraries, demonstrating its capability to handle synergistic preferences. However, balancing inherently conflicting objectives remains challenging. This highlights the need for future advancements, such as domain-specific fine-tuned LLMs to better resolve preference conflicts or multi-objective optimization techniques to systematically navigate trade-offs.

## I.2 Open Reasoning with Chinese Context

In this section, we quantitatively compare the reasoning capabilities of LLMs in the context of Chinese travel requirements. Given that many leading LLMs, such as GPT-4, are primarily trained in English corpora, it is essential to evaluate their performance in a Chinese travel planning context to better understand their reasoning abilities. We focus on three LLMs: GPT-4o, DeepSeek-V2.5, and Qwen2.5-7B, which are employed in the main experiments.

Specifically, we analyze the POI matching in the NL2DSL process with varying travel requirements from the synthesized quires and further provide the distribution of the results in Figure 13. The comparative analysis reveals significant disparities in reasoning capabilities across the three LLMs when handling Chinese travel-related queries. DeepSeek-V2.5 demonstrates robust performance in most categories, achieving high accuracy (Correct ≥ 93%) for attraction-names, attraction-types, restaurant-names, and hotel-features. However, its performance sharply declines in hotel-names (Correct: 67%, Missing: 33%), suggesting limited familiarity with Chinese hotel nomenclature or insufficient contextual grounding in this domain. This contrasts with GPT-4o, which excels in

hotel-names (Correct: 93%) and achieves perfect accuracy (Correct: 100%) for attraction-types, highlighting its superior cross-lingual transfer capabilities despite being primarily English-trained. Notably, GPT-4o maintains consistent performance across all categories (Correct  $\geq$  93%), underscoring its balanced reasoning proficiency in Chinese contexts. In stark contrast, Qwen2.5-7B exhibits critical weaknesses, particularly in attraction-names (Correct: 13%, Error: 43%), indicating severe limitations in entity recognition and syntactic coherence for Chinese proper nouns. The pronounced missing rates observed in Qwen2.5-7B (e.g., 43% for attraction-names and 23% for hotel-names) align with its constrained parameter size (7B), which likely impedes its ability to internalize diverse travel requirements or align them with sandbox’s POI information.

We further conduct the analysis and provide the results on human queries in Figure 14. The evaluation of human queries reveals critical limitations in LLMs’ practical reasoning capabilities that synthetic data fails to expose. DeepSeek-V2.5’s accuracy plummets in hotel-feature (Correct: 40% vs. 93% in synthetic data), indicating severe degradation when handling ambiguous or culturally nuanced requirements (e.g., interpreting subjective descriptors like “luxury” or “traditional courtyard-style” in Chinese contexts). GPT-4o similarly exhibits instability, with significant declines in restaurant-types (Correct: 37% vs. 97% in synthetic data) and attractions-type (Correct: 69% vs. 100%), suggesting that its cross-lingual transfer mechanisms falter when confronted with real-world linguistic variability (e.g., colloquial phrasing or dialect influences). This analysis underscores the necessity of introducing human queries into benchmarks when evaluating travel planning, as they reveal critical gaps in open language reasoning for deploying LLMs in real-world travel assistants.

### 1.3 Analysis of Pure-LLM Methods

Pure LLM-based methods have demonstrated significant shortcomings in constraint satisfaction, as evidenced by their near-zero success rates in benchmarks like TravelPlanner. We also attempt the multi-round refinement methods like Reflexion. While theoretically promising, it is still impractical in our context. In preliminary evaluations, Reflexion not only failed to achieve improvements in constraint satisfaction (consistent 0% FPR) but also incurred prohibitive computational costs due to its reliance on iterative token-heavy interactions. This rendered large-scale evaluation infeasible given our resource constraints. In light of their current limitations in constraint satisfaction, NeSy frameworks remain the effective pathway for real-world travel planning. Therefore, in the main body of this work, we mainly analyze the NeSy method.

In this section, we further summarize the key failure modes of pure-LLM-based methods observed in our experiments:

(1) **Incorrect API Calls:** LLMs frequently generate invalid or hallucinated API calls, leading to cascading errors in downstream planning. For instance, models may query non-existent APIs (e.g., `city_transport.select` instead of `inter_city_transport.select`) or misuse parameters (e.g., filtering attractions by an unsupported feature like “bus”). Such errors exhaust API call limits and prevent agents from retrieving essential information.

(2) **Repetitive Output Loops** In iterative planning frameworks like ReAct, LLMs often enter infinite loops when resolving constraints. For example, an agent might repeatedly query transportation details for all candidate attractions, even after selecting one, due to a failure to update its internal state. This behavior mimics the “hallucination loops” reported in TravelPlanner paper.

(3) **Reasoning-Action Inconsistency.** In ReAct framework, the model first reasons and then takes an action. However, the reasoning and the action are not always consistent. For example, the model may reason that the user wants to book a flight, but then take an action to check the information of trains. Another example is that the model may detect that the expenses exceed the budget but does not respond to this and ultimately generates a plan that exceeds the budget.

(4) **Critical Information Missing.** Even when intermediate steps (e.g., API responses) are logged in a “notebook,” LLMs frequently omit essential details when synthesizing final plans. A recurring failure is neglecting return transportation (e.g., omitting the train from Shanghai back to Beijing), which violates feasibility constraints.

Figure 15 provides the fail examples of ReAct (one-shot) with DeepSeek, which outperforms other pure-LLM-based methods in the main experiments.

815 These limitations underscore the inadequacy of pure-LLM-based approaches for deployment in  
816 long-horizon and constraint-rich domains like travel planning.

## 817 J Statements about Scientific Artifacts

818 The ChinaTravel benchmark is designed to facilitate research in natural language processing and  
819 artificial intelligence, specifically for travel planning tasks. ChinaTravel includes a travel sandbox,  
820 user queries, and an evaluation framework intended for non-commercial, academic research purposes.

821 **Availability.** We will publicly release the ChinaTravel benchmark upon publication to facilitate  
822 community research. We look forward to broader adoption and extension of this benchmark.

823 **Licenses.** The ChinaTravel benchmark and its associated datasets are licensed under the Creative  
824 Commons Attribution-NonCommercial 4.0 International (CC-BY-NC 4.0) license. This license  
825 allows for the free use, distribution, and reproduction of the benchmark in any medium, provided that  
826 appropriate credit is given to the original authors and the source of the data is acknowledged, and that  
827 the use is for non-commercial purposes only.

828 **Data anonymization and offensive content.** We anonymized the human queries during collection  
829 and instructed participants to avoid including sensitive information. We removed queries containing  
830 offensive content during the data cleaning process.

## 831 K Statements about Human Participants

832 We recruited over 250 volunteers through a structured questionnaire to collect authentic Chinese travel  
833 requirements. Participants were informed about the public use of their data and instructed to avoid  
834 including sensitive personal information. During data cleaning, offensive content and identifiable  
835 details were removed. While no explicit ethics board approval is mentioned, we ensured compliance  
836 with anonymization practices and obtained participant consent for data inclusion. The final dataset  
837 contains 154 human-derived queries reflecting diverse real-world travel needs.

### 838 K.1 Instructions Given To Participants

839 To gather the authentic travel requirements, we collected data through a carefully designed question-  
840 naire. We provided the following instruction information to the participants:

- 841 1. The specific constraints the agent can handle and the corresponding details, including the  
842 types and specific names of attractions, restaurants, and hotels; requirements for intercity  
843 transportation (airplane or train) and urban transportation (walk, taxi or subway); as well as  
844 budget limitations for overall expenses or specific activities (such as accommodation and  
845 intercity transportation).
- 846 2. The necessary information should be provided in the query, including the departure and  
847 destination cities of the trip, the number of travel days and constraint information.
- 848 3. A detailed example with the query and travel planning response.

849 Fig. 16 and Fig. 17 respectively show the questionnaire and its translated version.

### 850 K.2 Recruitment And Payment

851 For the collection of Human-154, we recruited a total of 250 student volunteers to provide authentic  
852 Chinese travel requirements. The participants included 121 undergraduate students, 86 master’s  
853 students, and 43 doctoral students. The task of understanding the query background and providing  
854 travel requirements was estimated to take 1-2 minutes per participant. Given the simplicity of the task  
855 and the fact that it did not require extensive professional background or expertise, we compensated  
856 each participant with 1 yuan. This compensation was deemed adequate considering the nature of the  
857 task and the time required to complete it. The payment was determined based on the estimated time



858 and the straightforward nature of the natural language requirements, ensuring a fair and reasonable  
859 reward for the participants.

860 For Human-1000, we partnered with WJX (a professional survey platform) to scale data collection.  
861 Each valid query was incentivized with 6 CNY. After WJX’s initial screening, our team rigorously  
862 annotated responses, filtering invalid entries (e.g., nonsensical inputs). It finally yielded 1,000  
863 high-quality queries meeting DSL annotation standards, ensuring both diversity and alignment with  
864 real-world planning scenarios.

### 865 **K.3 Data Consent**

866 When collecting the data, we clearly informed the participants about the usage of the data and the  
867 potential irreversible risks of it becoming part of a public dataset. We did not track the ID information  
868 of the questionnaire respondents. Additionally, we reminded participants not to include any sensitive  
869 personal information in the questionnaire responses. During the data cleaning process, we directly  
870 removed queries containing offensive content and filtered out sensitive identity information.

### 871 **K.4 Characteristics of Annotators**

872 Our data collection process solely involved travel requirements and did not include any protected  
873 information, such as sexual orientation or political views as defined under the General Data Protection  
874 Regulation (GDPR). All data were collected from native Chinese speakers to ensure that the travel  
875 requirements fully align with the context and nuances of the Chinese language. This approach was  
876 taken to accurately capture the needs and preferences of the target population, which is primarily  
877 composed of Chinese-speaking individuals. The annotators were recruited from a diverse range of  
878 academic backgrounds, including undergraduate, master’s, and doctoral students, to provide a broad  
879 and representative set of travel requirements.

### 880 **K.5 DSL Annotation for Human Data**

881 The annotation process for the human data involved four stages to ensure the accuracy and validity of  
882 the Domain-Specific Language (DSL) annotations: (1) Initial DSL Version Generation: GPT-4o was  
883 utilized to provide the initial version of the DSL annotations for the human data. This step aimed  
884 to leverage the language model’s capabilities to generate a baseline for further refinement. (2) Data  
885 Annotation Team Revision: A team of five data annotators was responsible for reviewing and revising  
886 the DSL annotations. The team members divided the workload and made necessary corrections to  
887 the DSL annotations to ensure their accuracy and relevance to the travel requirements. (3) Primary  
888 Developer Verification and Correction: Three of the main developers of the benchmark conducted  
889 a thorough review of all the DSL annotations. They verified the correctness of the annotations and  
890 made revisions as needed. This stage also involved the exclusion of any invalid queries that could not  
891 be verified within the sandbox environment. (4) Final Verification by Primary Developers: The same  
892 three main developers performed a final check on all the DSL annotations. This step ensured that the  
893 annotations were accurate, consistent, and met the required standards for the benchmark.

894 Throughout the annotation process, the focus was on ensuring that the DSL annotations accurately  
895 captured the travel requirements and were valid within the context of the ChinaTravel benchmark’s  
896 sandbox environment. The annotation process for human data required a deep understanding of the  
897 ChinaTravel DSL and involved joint debugging and verification with the sandbox information. This  
898 significantly limited the size of the annotation team, as only a limited number of annotators had the  
899 necessary expertise and familiarity with both the DSL and the sandbox environment. Additionally,  
900 the process was time-consuming and required meticulous attention to detail, further constraining  
901 the rate at which the human dataset could grow. Despite these challenges, the rigorous annotation  
902 process ensured the quality and reliability of the human data, which is crucial for the evaluation and  
903 development of language agents in real-world travel planning.

## L TTG

### L.1 Constraints Formulation

TTG [12] models the travel planning problem as a MILP (Mixed-Integer Linear Programming) problem. We adapt their formulation into ChinaTravel for solver-based optimization and the specific parameters, variable and constraint settings can be found in Tab. 151617.

### L.2 Experiment

Although TTG performs very well on Travelplanner, the solver takes slightly more than 1 second on average to complete the computation. On the ChinaTravel benchmark, the rapid growth of constraints in TTG becomes computationally prohibitive. If we use the full sandbox, the average number of constraints will exceed **10B** (For detailed calculations of variable sizes and the number of constraints, please refer to Tab. 1819). Therefore, we only include 22 POIs (2 hotels, 10 attractions, 5 restaurants, 5 stations, 100 intercity transports each for arrivals and departures) and use one hour as a time step. We use LLMs to select them from sandbox to ensure sufficient flexibility in handling different queries. Nonetheless, its constraint scale still reaches  $320k \times \text{days}$  and the number of variables also reaches  $36k \times \text{days}$ . In comparison, the commonly used benchmark for evaluating MILP solvers, MIPLIB 2017 [6], contains only 10 instances with more than 320k constraints and about 60 instances with over 36k variables (out of a total of 1065 instances).

In our main experiments, using the SCIP solver from the PuLP package, TTG was allocated a relaxed 15-minute search limitation. However, this configuration yielded only 18% valid solutions on easy-subset instances, with the false positive rate (FPR) further reduced to 8% due to the solver’s pruning heuristics. Fig. 10(a) illustrates the solution time of TTG on 1- to 3-day itineraries. Within the time limit, solutions were found for merely 23% of two-day and 6% of three-day itineraries.

Parameter	Meaning
<i>hotelNum</i>	Number of hotels
<i>attrNum</i>	Number of attractions
<i>restNum</i>	Number of restaurants
<i>transNum</i>	Number of transport modes
<i>stationNum</i>	Number of stations
<i>goNum</i>	Number of arriving trains/buses
<i>backNum</i>	Number of departing trains/buses
<i>timeStep</i>	Number of time steps
$locNum = hotelNum + attrNum + restNum$	Total number of POI locations except stations
$totalNum = locNum + stationNum$	Total number of all locations including stations

Table 15: Definition of parameters used in TTG

Variable	Meaning
$u[idx][t]$	The traveler is at location $idx$ at time $t$
$event[t]$	The traveler's location changes at time $t$
$hotel[idx][d]$	Number of times the traveler visits hotel $idx$ on day $(d + 1)$
$attr[idx]$	Number of times the traveler visits attraction $idx$
$rest[idx][meal]$	Number of times the traveler visits restaurant $idx$ at meal $meal$
$z_{hotel}, z_{attr}, z_{rest}, \delta$	Auxiliary variables
$needEat[m]$	Whether the traveler needs to eat meal $m$ (during intercity travel)
$check[idx][t]$	Whether the attraction $idx$ is open at time $t$
$y[i, j, tr, t]$	The solution, a matrix of shape $totalNum \times totalNum \times transNum \times timeStep$

Table 16: Variables used in TTG

An Example of Prompts for Data Generation
<pre> # You are a user who wants to ask an AI agent to help you   plan a trip. Please construct some natural language   inquiries based on the following example and provide   the corresponding logical constraint expressions. Note   that "tickets" and "people_number" are the same. # Example: # JSON: # {} # Use the following restaurants. # Restaurant name: {} # This means that "restaurant_names" should include this   restaurant. # The dining options may not always be exactly as   described by the provided features; synonyms can be   used. For example, if the hotel's feature is a pool,   you could ask naturally in language like "I want to   swim in the hotel pool." # Now, your departure location is {}, and your destination   is {}. The number of people is {}, and the number of   days is {}. # Now please provide a JSON inquiry. # JSON: </pre>

Figure 7: An example of prompts for data generation. This example is about restaurant\_name. By replacing this with other constraints or combining multiple constraints, we can generate data with different levels of difficulty based on different constraints.

## Examples of Generated Data

### Example 1

```
{
  "start_city": "杭州",
  "target_city": "上海",
  "hard_logic": [
    "days==2",
    "people_number==1",
    "tickets==1",
    "{ '本帮菜' } i= food.type"
  ],
  "nature_language": "当前位置杭州。我一个人想去上海玩2天，想尝试当地的特色菜，请给我一个旅行规划。"
}
```

### Example 2

```
{
  "start_city": "深圳",
  "target_city": "北京",
  "hard_logic": [
    "days==2",
    "people_number==3",
    "intercity_transport=={'airplane'}",
    "tickets==3",
    "rooms==3",
    "room_type==1"
  ],
  "nature_language": "当前位置深圳。我们三个人计划去北京玩两天，选择飞机出行，开三间大床房。请给我一个旅行规划。"
}
```

### Example 3

```
{
  "start_city": "重庆",
  "target_city": "苏州",
  "hard_logic": [
    "days==3",
    "people_number==3",
    "cost_i=7300",
    "{ '日本料理' } i= food.type",
    "intercity_transport=={'train'}",
    "tickets==3",
    "rooms==2",
    "room_type==2"
  ],
  "nature_language": "当前位置重庆。我们三个人计划去苏州玩三天，选择火车出行，想吃日本料理，预算7300元，开两间双床房。请给我一个旅行规划。"
}
```

Figure 8: Examples of Generated Data

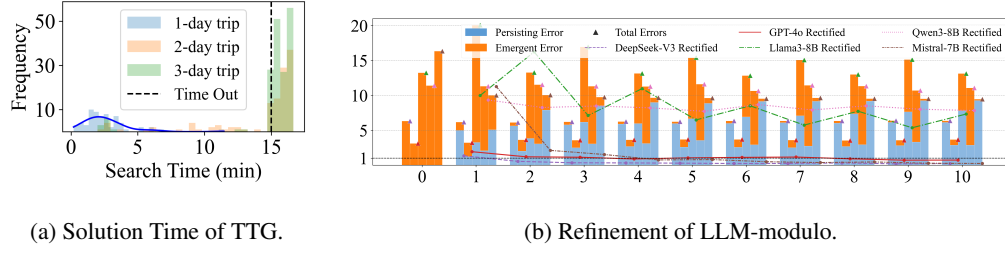


Figure 10: (a) The high computational complexity of TTG renders it infeasible for real-world multi-day itineraries. (b) LLM-modulo’s error correction declines during iteration, causing emergent errors.

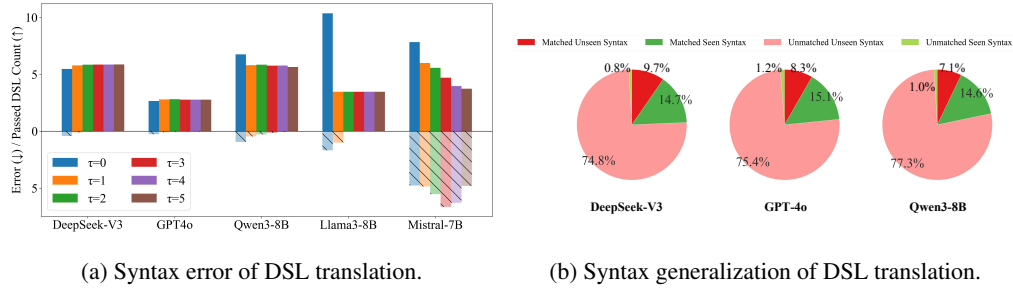


Figure 11: Challenges in NL2DSL translation.

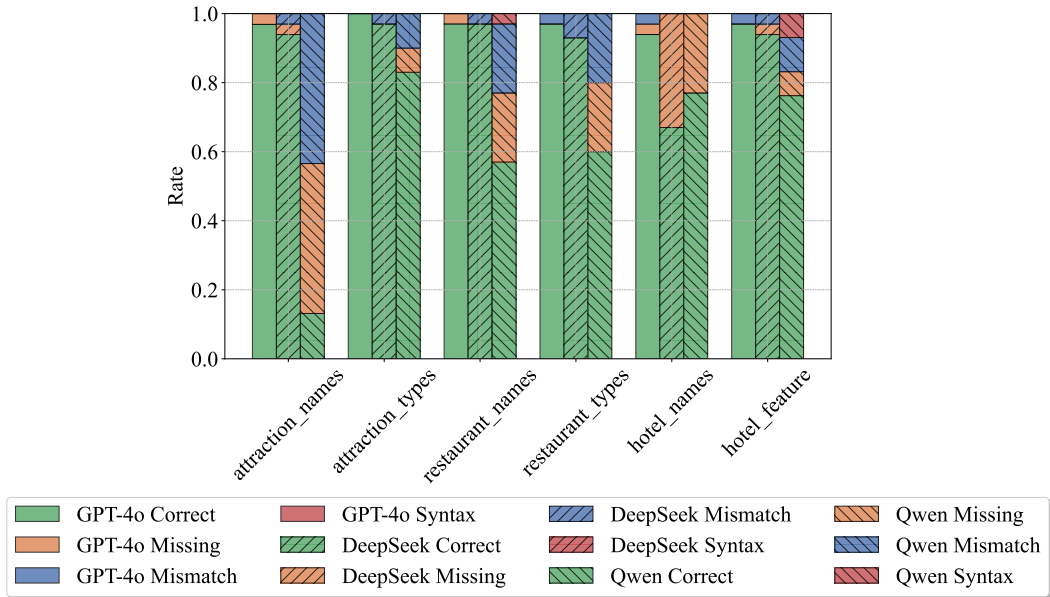


Figure 13: Results Distribution on Synthesized Quieres

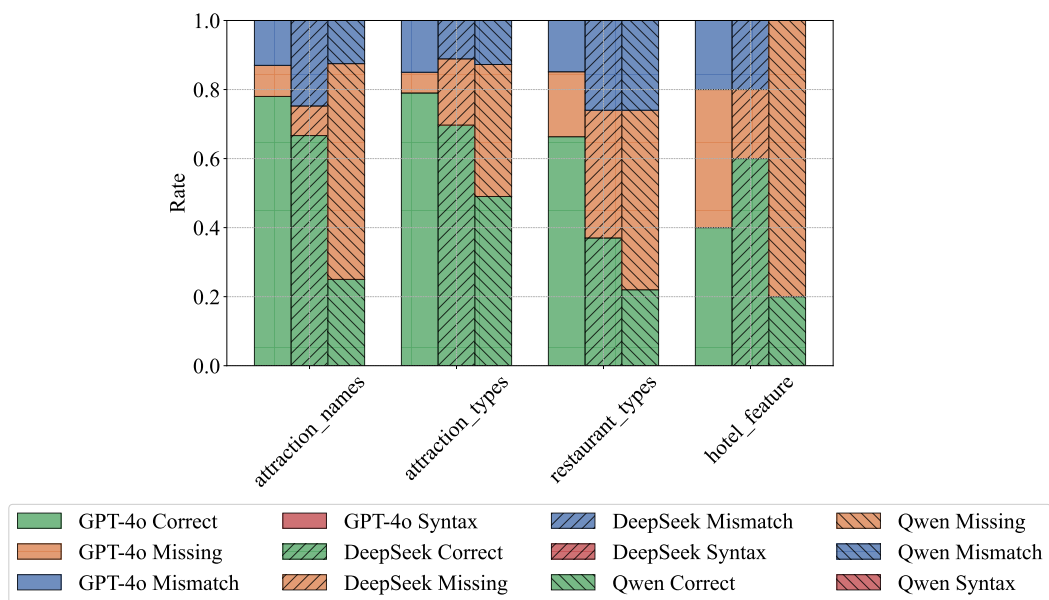


Figure 14: Results Distribution on Human Quires

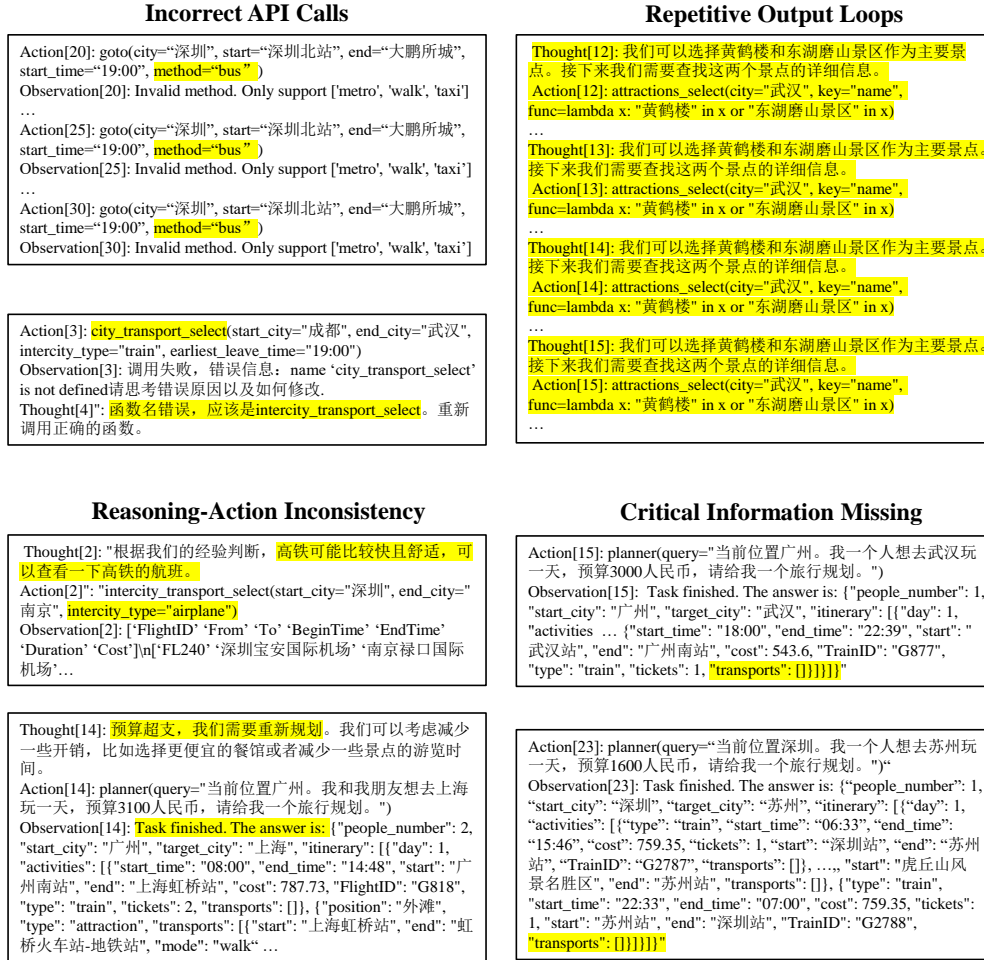


Figure 15: Fail case studies of React-one-shot DeepSeek Method.

开放旅行规划问题搜集

本问卷旨在构建一个开放环境下的旅行规划数据集，以便于相关研究的开展。由于填写的问题将作为公开数据集的一部分，存在无法撤销的风险；请勿在填写内容中包含任何敏感的个人信息，感谢大家的参与！

1. 出发城市：\_\_\_\_\_（从北京、南京、上海、杭州、深圳、武汉、广州、成都、重庆、苏州中选择）

2. 目标旅游城市：\_\_\_\_\_（从北京、南京、上海、杭州、深圳、武汉、广州、成都、重庆、苏州中选择）

3. 旅行人数：\_\_\_\_\_（1-5）

4. 旅行天数：\_\_\_\_\_（1-5）

您作为用户可以向智能代理发起查询请求。查询内容可以包括对景点、餐饮、住宿、跨城交通(如火车、飞机)以及城内交通(如地铁、步行、出租车)的具体要求。同时，您也可以提供个人偏好。请确保查询中包含以下三个信息:目标城市、人数和天数，并确保这些信息相互匹配。智能代理将根据您的请求提供一个旅行规划结果，包括这几天的交通安排、住宿地点、推荐的景点及餐饮建议。

用户问题的例子：

当前位置苏州。我一个人想去南京玩 2 天，预算 3000 人民币，往返都坐高铁，请给我一个旅行规划。

智能代理回复的例子：

起点:苏州

目的地:南京

交通:苏州北站 -> 南京南站

列车:G4，07:24->08:15

费用:122.9 元

车票:1 张

游览:玄武湖景区

交通:地铁(南京南站 ->南京林业大学·新庄)，步行 3 分钟 +地铁 23 分钟+步行 8 分钟

费用:4 元

游览时间:08:50->10:00

门票:0 元

.....

午餐:南京金鹰国际酒店·满园春中餐厅

费用:188 元

时间:12:10 ->13:10

住宿:桔子水晶南京玄武湖酒店

房型:大床房，1 间

费用:370 元

返回:南京南站 > 苏州站

列车:G7220，20:09->21:23

费用:122.9 元

车票:1 张

我们将用户问题分为不同难度级别进行分类，以下是每个级别的描述

低级:涉及一般性问题，不包含个性化需求。

中级:包含一定程度的个性化需求，通常涉及到食宿交通等方面。

高级:涉及更复杂、更具体的需求，如时间要求、特定地点或活动的安排等。

以下是不同难度级别下的用户问题示例：

低级:我想知道去上海玩 2 天的行程规划，从杭州出发。

中级:我想独自一人前往南京穷游，计划在那里待 3 天左右。我对历史文化很感兴趣，希望能深度游览一些古迹。

高级:我们三人后天需要前往北京玩 2 天。第二天晚上十点前需要从北京站返回。想在第一天去故宫，第二天去天坛，请给我一个旅行规划

5. 请给出用户问题：\_\_\_\_\_

Figure 16: Questionnaire



### Open Travel Planning Data Collection Questionnaire

This questionnaire aims to construct a dataset for travel planning in an open environment to facilitate relevant research. Since the responses will be part of a public dataset and cannot be revoked, please do not include any sensitive personal information in your responses. Thank you for your participation!

- Departure City:\_\_\_\_\_ (Choose from Beijing, Nanjing, Shanghai, Hangzhou, Shenzhen, Wuhan, Guangzhou, Chengdu, Chongqing, Suzhou)
- Destination City: \_\_\_\_\_ (Choose from Beijing, Nanjing, Shanghai, Hangzhou, Shenzhen, Wuhan, Guangzhou, Chengdu, Chongqing, Suzhou)
- Number of Travelers:\_\_\_\_\_ (1-5)
- Number of Travel Days:\_\_\_\_\_ (1-5)

As a user, you can submit queries to the intelligent agent. Your query may include specific requirements for attractions, dining, accommodation, intercity transportation (e.g., train, plane), and intra-city transportation (e.g., subway, walking, taxi). You may also provide personal preferences. Please ensure that your query includes the following three pieces of information: the destination city, the number of travelers, and the number of travel days, and make sure they are consistent. The intelligent agent will generate a travel plan based on your request, covering transportation arrangements, accommodation, recommended attractions, and dining suggestions.

Example User Query:  
 "My current location is Suzhou. I want to travel alone to Nanjing for 2 days with a budget of 3,000 RMB, taking the high-speed train for both departure and return. Please provide a travel plan."

Example Response from the Intelligent Agent:

Departure: Suzhou  
 Destination: Nanjing  
 Transportation: Suzhou North Station → Nanjing South Station  
 Train: G4, 07:24 → 08:15  
 Cost: 122.9 RMB  
 Tickets: 1  
 Attraction: Xuanwu Lake Scenic Area  
 Transportation: Subway (Nanjing South Station → Nanjing Forestry University-Xinzhuang)  
 Route: Walk 3 minutes → Subway 23 minutes → Walk 8 minutes  
 Cost: 4 RMB  
 Visit Time: 08:50 → 10:00  
 Admission: 0 RMB  
 ...  
 Lunch: Nanjing Jinling Hotel · Man Yuan Chun Chinese Restaurant  
 Cost: 188 RMB  
 Time: 12:10 → 13:10  
 Accommodation: Crystal Orange Hotel Nanjing Xuanwu Lake  
 Room Type: Queen Room, 1 room  
 Cost: 370 RMB  
 Return: Nanjing South Station → Suzhou Station  
 Train: G7220, 20:09 → 21:23  
 Cost: 122.9 RMB  
 Tickets: 1

Classification of User Queries by Difficulty Level  
 We categorize user queries into different difficulty levels as follows:

Easy Level: General inquiries without personalized requirements.  
 Medium Level: Includes some degree of personalization, usually involving food, lodging, or transportation.  
 Hard Level: Involves more complex and specific needs, such as time constraints, particular locations, or planned activities.

Examples of User Queries at Different Difficulty Levels:  
 Basic Level: "I want to know the itinerary for a 2-day trip to Shanghai from Hangzhou."  
 Intermediate Level: "I plan to travel alone to Nanjing on a budget and stay for about three days. I'm interested in history and culture and would like to explore historical sites in depth."  
 Advanced Level: "Three of us need to travel to Beijing the day after tomorrow for a 2-day trip. We need to return from Beijing Railway Station before 10 PM on the second day. We want to visit the Forbidden City on the first day and the Temple of Heaven on the second day. Please provide a travel plan."

- Please provide a user query:\_\_\_\_\_

Figure 17: The translated version of the questionnaire

ChinaTravel	TravelPlanner
<p>当前位置武汉。我一个人想去苏州玩一天，预算 1400 人民币，请给我一个旅行规划。</p> <p>Current location: Wuhan. I want to visit Suzhou for a day by myself with a budget of 1,400 RMB. Please provide me with a travel plan.</p>	<p>Please help me plan a trip from St. Petersburg to Rockford spanning 3 days from March 16th to March 18th, 2022. The travel should be planned for a single person with a budget of \$1,700.</p>
<p>当前位置南京。我一个人想去重庆玩 3 天，喜欢吃甜食面包啥的，请给我一个旅行规划。</p> <p>Current location: Nanjing. I want to travel to Chongqing alone for 3 days. I like sweet foods and bread. Please provide me with a travel plan.</p>	<p>Please design a travel plan departing from Las Vegas and heading to Stockton for 3 days, from March 3rd to March 5th, 2022, for one person, with a budget of \$1,400.</p>
<p>当前位置重庆。我和朋友两个人想去武汉玩 3 天，想尝试当地菜，请给我们一个旅行规划。</p> <p>Current location: Chongqing. My friend and I want to visit Wuhan for 3 days and try the local cuisine. Could you please provide us with a travel plan?</p>	<p>Craft a travel plan for me to depart from New Orleans and head to Louisville for 3 days, from March 12th to March 14th, 2022. I will be travelling alone with a budget of \$1,900.</p>
<p>当前位置成都。我们三个人想去深圳玩 2 天，想去历史感比较重的景点，请给我们一个旅行规划。</p> <p>Current location: Chengdu. The three of us want to visit Shenzhen for 2 days and are interested in historical sites. Could you please provide us with a travel itinerary?</p>	<p>Could you aid in curating a 5-day travel plan for one person beginning in Denver and planning to visit 2 cities in Washington from March 23rd to March 27th, 2022? The budget for this trip is now set at \$4,200.</p>
<p>当前位置深圳。我和朋友两个人想去上海玩 3 天，想去海洋水族馆，请给我们一个旅行规划。</p> <p>Current location: Shenzhen. My friend and I want to visit Shanghai for 3 days and we would like to go to the Ocean Aquarium. Could you please provide us with a travel plan?</p>	<p>Could you assist in crafting a travel itinerary for a 5-day, single-person trip departing from Orlando and touring 2 cities in Texas? The travel dates should range from March 10th to March 14th, 2022, and the entire travel budget is \$3,100.</p>
<p>当前位置成都。我和朋友两个人想去上海玩 3 天，住一间双床房，期间可能要开会，酒店最好能提供个开会的地点，请给我一个旅行规划。</p> <p>Current location: Chengdu. My friend and I want to visit Shanghai for 3 days. We need a twin room, and we might need a meeting space during our stay. Please provide me with a travel plan.</p>	<p>Could you help me arrange a 7-day solo travel itinerary from Kona to California with a budget of \$5,800, intending to visit 3 distinct cities in California from March 7th to March 13th, 2022?</p>
<p>我目前在南京，计划和两个朋友一起去上海玩两天，选择原舍·在水一方度假酒店，请帮我们规划一个旅行方案。</p> <p>I am currently in Nanjing and plan to travel to Shanghai with two friends for two days. We have chosen the YuanShe · Zai Shui Yi Fang Resort Hotel. Please help us plan a travel itinerary.</p>	<p>Please help me craft a 7-day travel plan. I'm planning on leaving from Punta Gorda and exploring 3 different cities in Wisconsin from March 16th to March 22nd, 2022. The budget for this trip is set at \$5,700.</p>
<p>当前位置北京。我和三个朋友计划去成都玩两天，选择火车出行，市内交通方式为地铁。请给我一个旅行规划。</p> <p>Current location: Beijing. My three friends and I are planning to visit Chengdu for two days. We have chosen to travel by train and use subway for city transportation. Please provide me with a travel itinerary.</p>	<p>Could you help me create a 7-day travel plan starting on March 18th, 2022, and ending on March 24th, 2022? The trip will start in Washington and I would like to visit 3 cities in Minnesota. This trip is for one person with a budget of \$7,200.</p>

Figure 18: Examples of easy-level queries from ChinaTravel and TravelPlanner.

ChinaTravel	TravelPlanner
<p>当前位置武汉。我两个人想去苏州玩 2 天，预算 4000 人民币，坐火车去，住一间大床房，想去虎丘山风景名胜区的自然风光，请给我一个旅行规划。</p> <p>Current location: Wuhan. Two of us want to visit Suzhou for 2 days with a budget of 4000 RMB. We plan to take the train and stay in a room with a king-size bed. We would like to visit natural attractions like Tiger Hill Scenic Area. Please provide a travel itinerary.</p>	<p>Could you please arrange a 3-day trip for two, starting in Sacramento and heading to Atlanta, from March 14th to March 16th, 2022. The budget for this trip is \$4,700, and we require accommodations where parties are allowed.</p>
<p>当前位置广州。我两个人想去成都玩 3 天，预算 9000 人民币，坐火车往返，住一间大床房，麻烦给我一个旅行规划。</p> <p>Current location: Guangzhou. Two of us want to visit Chengdu for 3 days with a budget of 9,000 RMB. We plan to travel round-trip by train and stay in a room with a double bed. Could you please provide a travel itinerary for us?</p>	<p>Could you please design a 3-day travel plan for a group of 5, departing from Manchester and heading to Charlotte, from March 29th to March 31st, 2022? Our budget is set at \$4,800 and we would prefer to have entire rooms for our accommodations.</p>
<p>当前位置广州。我和我的两个朋友想去深圳玩两天，预算 2100 人民币，住两间双床房，坐地铁游玩，想吃海鲜，想去深圳欢乐谷玩。Current location: Guangzhou. My two friends and I want to go to Shenzhen for two days. Our budget is 2,100 RMB. We plan to stay in two twin-bed rooms, travel around by metro, eat seafood, and visit Shenzhen Happy Valley.</p>	<p>Could you tailor a 5-day travel plan for two people, departing from Knoxville and visiting 2 cities in Florida from March 20 to March 24, 2022? Our budget is set at \$3,900. We'd love to explore local Chinese and Mediterranean cuisines during our stay.</p>
<p>当前位置武汉。我两个人想去杭州玩 3 天，预算 7000 人民币，坐飞机往返，住一间大床房，麻烦给我一个旅行规划。</p> <p>Current location: Wuhan. Two of us want to visit Hangzhou for 3 days with a budget of 7,000 RMB. We plan to travel by plane round-trip and stay in a room with a large bed. Could you please provide a travel plan for us?</p>	<p>Could you help create a 7-day travel plan for a group of 3, departing from Greensboro and touring 3 different cities in Georgia from March 10th to March 16th, 2022? We have a new budget of \$4,000 for this trip. We'd also appreciate if our accommodations have smoking areas.</p>
<p>当前位置杭州。我两个人想去苏州玩 2 天，预算 3500 人民币，住一间大床房，想去看看拙政园这样的园林景观，请给我一个旅行规划。</p> <p>Current location: Hangzhou. Two of us want to visit Suzhou for 2 days with a budget of 3,500 RMB. We would like to stay in a room with a large bed and visit garden attractions like the Humble Administrator's Garden. Please provide a travel plan.</p>	<p>Could you help create a 5-day travel itinerary for a group of 4, starting from New York and visiting 2 cities in Louisiana from March 15th to March 19th, 2022? We have a budget of \$12,300. Please note that we require accommodations where smoking is permissible.</p>
<p>当前位置北京。我两个人想去深圳玩 3 天，预算 7000 人民币，住一间大床房，坐飞机去，酒店最好有泳池，想去深圳欢乐谷看一下，请给我一个旅行规划。</p> <p>Current location: Beijing. Two of us want to visit Shenzhen for 3 days with a budget of 7,000 RMB. We would like to stay in a hotel with a king-size bed and preferably a swimming pool. We plan to fly there and would like to visit Shenzhen Happy Valley. Please provide a travel itinerary.</p>	<p>Can you provide me with a 5-day travel plan for 2 people, starting from Asheville and exploring 2 cities in New York from March 13th to March 17th, 2022? Our budget is set at \$4,700 and we would love to try local Mexican and Chinese cuisines during our trip.</p>

Figure 19: Examples of medium-level queries from ChinaTravel and TravelPlanner.

ChinaTravel	TravelPlanner
<p>[当前位置武汉,目标位置南京,旅行人数 2,旅行天数 4] 我和同学 2 人打算去南京玩 4 天, 预算 1500 (不包括车票住宿), 只是玩和吃饭, 请你帮忙规划。</p> <p>[Current location: Wuhan, Destination: Nanjing, Number of travelers: 2, Duration of travel: 4 days] My classmate and I are planning to visit Nanjing for 4 days. Our budget is 1500 (excluding transportation and accommodation), just for activities and meals. Please help us plan.</p>	<p>Can you create a 5-day itinerary for a group of 7 people traveling from Richmond to two cities in Florida between March 9th and 13th, 2022? Our budget is \$8,500. We require accommodations that allow visitors and should ideally be entire rooms. In regards to dining options, we prefer French, American, Mediterranean, and Italian cuisines.</p>
<p>[当前位置南京,目标位置成都,旅行人数 3,旅行天数 5] 我们一家三口想去成都旅游一周, 主要想逛一些适合带小朋友的景点, 预算 8000 元, 然后品尝一些当地的美食。</p> <p>[Current location: Nanjing, Destination: Chengdu, Number of travelers: 3, Travel days: 5] Our family of three wants to travel to Chengdu for a week. We mainly want to visit attractions suitable for children, with a budget of 8,000 yuan, and also taste some local delicacies.</p>	<p>Could you help design a travel plan for two people leaving from Houston to Pensacola for 3 days, from March 6th to March 8th, 2022? Our budget is set at \$1,400 for this trip and we require our accommodations to be visitor-friendly. We would like to have options to dine at Indian, American, Chinese, and Italian restaurants. We also prefer not to self-drive during the trip.</p>
<p>[当前位置广州,目标位置深圳,旅行人数 3,旅行天数 2] 我们一行三人要从广州去到深圳玩两天, 想去繁华的街区逛逛, 尽可能减少麻烦的交通, 总消费尽可能少。</p> <p>[Current location: Guangzhou, Destination: Shenzhen, Number of travelers: 3, Number of travel days: 2] Our group of three plans to travel from Guangzhou to Shenzhen for two days. We want to explore bustling neighborhoods, minimize inconvenient transportation, and keep the total expenses as low as possible.</p>	<p>Could you help create a 3-day travel plan for two people? We're traveling from West Palm Beach to White Plains, visiting only one city from March 5th to March 7th, 2022. We have a budget of \$2,600. For our accommodations, we'd like rooms that are not shared. We are not planning on self-driving and will be reliant on public transportation. Cuisines we are interested in trying include Mexican, Chinese, Mediterranean, and American.</p>
<p>[当前位置苏州,目标位置杭州,旅行人数 4,旅行天数 2] 我想 4 个人去杭州 2 天进行历史文化遗址的考察顺带玩一下。</p> <p>[Current location: Suzhou, Destination: Hangzhou, Number of travelers: 4, Duration of travel: 2 days] I would like 4 people to go to Hangzhou for 2 days to explore historical and cultural sites and have some fun along the way.</p>	<p>Could you generate a 3-day travel plan for a group of 3 people, departing from Bangor and visiting Washington from March 21st to March 23rd, 2022? Our budget is set at \$3,100. We require accommodations that are pet-friendly and we would prefer to have entire rooms to ourselves. We do not plan on self-driving for this trip</p>
<p>[当前位置上海,目标位置北京,旅行人数 1,旅行天数 3] 我要从上海出发, 到北京玩三天, 希望看一些名胜古迹, 吃一些当地特色, 预算充分。</p> <p>[Current location: Shanghai, Destination: Beijing, Number of travelers: 1, Number of travel days: 3] I want to depart from Shanghai and spend three days in Beijing. I hope to see some famous landmarks and try some local specialties, with a sufficient budget.</p>	<p>Could you help with creating a 5-day travel plan for 2 people, originating from Evansville and covering 2 cities in Texas from March 17th to March 21st, 2022? Our preferred accommodations are private rooms, and they must permit children under 10 since we will have them with us. Transportation should not involve any flights. The budget for this trip is set at \$2,800.</p>
<p>[当前位置北京,目标位置上海,旅行人数 2,旅行天数 3] 我和朋友计划用三天的时间从北京到上海玩, 计划坐飞机来回, 偏红色旅游线路。</p> <p>[Current location: Beijing, Destination: Shanghai, Number of travelers: 2, Number of travel days: 3] My friend and I are planning to spend three days traveling from Beijing to Shanghai. We plan to fly round trip and prefer a red-themed travel route.</p>	<p>Can you assist in creating a travel itinerary for a group of 4, starting in Seattle and visiting 3 unique cities across Texas? This trip will span over 7 days from March 10th through March 16th, 2022. We have a budget of \$11,000. Regarding our accommodations, we would like to rent entire rooms, and it's important that our lodgings allow parties. As for transportation, we do not plan to drive ourselves around.</p>

Figure 20: Examples of human/hard level queries from ChinaTravel and TravelPlanner.

### Prompts for POI recommendation

```
NEXT_POI_TYPE_INSTRUCTION = """
You are a travel planning assistant.
The user's requirements are: {}.
Current travel plans are: {}.
Today is {}, current time is {}, current location is
    {}, and POI_type_list is {}.
Select the next POI type based on the user's needs and
    the current itinerary.
Please answer in the following format.
Thought: [Your reason]
Type: [type in POI_type_list]
"""
```

Figure 21: Prompts for next-POI-type recommendation

### Prompts for restaurants recommendation

```
RESTAURANT_RANKING_INSTRUCTION = """
    You are a travel planning assistant.
    The user's requirements are: {user_requirements}.
    The restaurant info is:
    {restaurant_info}
    The past cost for intercity transportation and hotel
    accommodations is: {past_cost}.

    Your task is to select and rank restaurants based on
    the user's needs and the provided restaurant
    information. Consider the following factors:
    1. Restaurant name
    2. Cuisine type
    3. Price range
    4. Recommended food

    Additionally, keep in mind that the user's budget is
    allocated across multiple expenses, including
    intercity transportation and hotel accommodations.
    Ensure that the restaurant recommendations fit
    within the remaining budget constraints after
    accounting for the past cost.
    Note that the price range provided for each restaurant
    is the average cost per person per meal, the
    remaining budget must cover the cost of three
    meals per day for {days} days.

    For each day, recommend at least 6 restaurants,
    combining restaurants for all days together.

    Your response should follow this format:

    Thought: [Your reasoning for ranking the restaurants]
    RestaurantNameList: [List of restaurant names ranked
    by preference, formatted as a Python list]
    """
```

Figure 22: Prompts for restaurant recommendation

### Prompts for attractions recommendation

```
ATTRACTION_RANKING_INSTRUCTION = """
    You are a travel planning assistant.
    The user's requirements are: {user_requirements}.
    The attraction info is:
    {attraction_info}
    The past cost for intercity transportation and hotel
    accommodations is: {past_cost}.

    Your task is to select and rank attractions based on
    the user's needs and the provided attraction
    information. Consider the following factors:
    1. Attraction name
    2. Attraction type
    3. Location
    4. Recommended duration

    Additionally, keep in mind that the user's budget is
    allocated across multiple expenses, including
    intercity transportation and hotel accommodations.
    Ensure that the attraction recommendations fit
    within the remaining budget constraints after
    accounting for the past cost.

    For each day, recommend at least 8 attractions,
    combining attractions for all days together. To
    ensure a comprehensive list, consider a larger
    pool of candidates and prioritize diversity in
    attraction type and location.

    Your response should follow this format:

    Thought: [Your reasoning for ranking the attractions]
    AttractionNameList: [List of attraction names ranked
    by preference, formatted as a Python list]

    Example:
    Thought: Based on the user's preference for historical
    sites and natural attractions, the attractions
    are ranked as follows:
    AttractionNameList: ["Attraction1", "Attraction2",
    ...]
    """
```

Figure 23: Prompts for attraction recommendation

Constraint Type	Mathematical Formulation
<b>Spatio-temporal Constraints</b>	$\delta[\text{idx}][t] \geq u[\text{idx}][t+1] - u[\text{idx}][t]$ $\delta[\text{idx}][t] \geq u[\text{idx}][t] - u[\text{idx}][t+1]$ $\text{event}[t] = 0 \Rightarrow u[\text{idx}][t] = u[\text{idx}][t+1]$ $\text{event}[t] = 1 \Rightarrow \sum_{\text{idx}} \delta[\text{idx}][t] = 2$ $\sum_i u[i][t] = 1$
<b>Hotel Constraints</b>	$z_{\text{hotel}}[\text{idx}][t] = u[\text{idx}][t] \wedge \text{event}[t]$ $\text{hotel}[\text{idx}][d] = \sum_{t=d \cdot \text{stepPerDay}}^{(d+1) \cdot \text{stepPerDay}} z_{\text{hotel}}[\text{idx}][t]$ $\sum_{\text{idx}} \text{hotel}[\text{idx}][d] = 1$
<b>Attraction Constraints</b>	$z_{\text{attr}}[\text{idx}][t] = u[\text{idx}][t] \wedge \text{event}[t]$ $\text{attr}[\text{idx}] = \sum_t z_{\text{attr}}[\text{idx}][t]$ $\sum_{\text{idx}} \text{attr}[\text{idx}] \geq \text{min\_attr}$ $\text{check}[\text{idx}][t] = \text{False} \Rightarrow u[\text{idx}][t] = 0$
<b>Meal Necessity</b>	$\text{needEat}[m] = 1 \Rightarrow a[m] < T_{\text{dep}}$ $\text{needEat}[m] = 1 \Rightarrow b[m] > T_{\text{arr}}$
<b>Innecity Transport Constraints</b>	$y[(i, j, \text{tran}, t)] \leq u[i][t]$ $y[(i, j, \text{tran}, t)] \leq \text{event}[t]$ $y[(i, j, \text{tran}, t)] \leq u[\text{tran}][t+1]$ $y[(i, j, \text{tran}, t)] \leq u[\text{tran}][t+\delta]$ $y[(i, j, \text{tran}, t)] \leq \text{event}[t+\delta]$ $y[(i, j, \text{tran}, t)] \leq u[j][t+\delta+1]$
<b>Restaurant Constraints</b>	$z_{\text{rest}}[\text{idx}][t] = u[\text{idx}][t] \wedge \text{event}[t]$ $\text{rest}[\text{idx}][m] = \sum_{t=a[m]}^{b[m]} z_{\text{rest}}[\text{idx}][t]$ $\sum_{\text{idx}} \text{rest}[\text{idx}][m] \leq \text{needEat}[m]$ $\text{check}[\text{idx}][t] = \text{False} \Rightarrow u[\text{idx}][t] = 0$
<b>Intercity Travel Constraints</b>	$\sum_i \text{interGo}[i] = 1$ $\sum_i \text{interBack}[i] = 1$ $\text{interGo}[i] = 1 \Rightarrow u[\text{goStation}[i]][t] = 1$ $\text{interBack}[i] = 1 \Rightarrow u[\text{backStation}[i]][t] = 1$

Table 17: Constraints used in TTG



Variable	Dimension
$u[\text{idx}][t]$	$(totalNum + transNum) \times timeStep$
$\delta[\text{idx}][t]$	$(totalNum + transNum) \times (timeStep - 1)$
$event[t]$	$timeStep$
$hotel[\text{idx}][d]$	$hotelNum \times days$
$z_{hotel}[\text{idx}][t]$	$hotelNum \times timeStep$
$attr[\text{idx}]$	$attrNum$
$z_{attr}[\text{idx}][t]$	$attrNum \times timeStep$
$rest[\text{idx}][meal]$	$restNum \times 3 \times days$
$z_{rest}[\text{idx}][t]$	$restNum \times timeStep$
$y[(i, j, tr, t)]$	$totalNum \times totalNum \times transNum \times timeStep$
total	$days \times stepPerHour \times 36k$

Table 18: Variable sizes in TTG

Category	Estimated Size
Spatio-temporal constraints	$(totalNum + transNum) \times (4 \times timeStep + 3)$
Hotel constraints	$hotelNum \times (3 \times timeStep + days)$
Attraction constraints	$4 \times attrNum \times timeStep$
Restaurant constraints	$restNum \times (4 \times timeStep + days)$
Urban transport constraints	$7 \times totalNum^2 \times transNum \times timeStep + 4 \times totalNum \times timeStep$
Intercity transport constraints	$(goNum + backNum) \times timeStep$

Table 19: Number of constraints sizes in TTG