Mind the Gap to Trustworthy LLM Agents: A Systematic Evaluation on Constraint Satisfaction for Real-World Travel Planning

Bo-Wen Zhang[1,2*], Jin Ye[1,2*], Jie-Jing Shao[1*], Yu-Feng Li[1,3], Lan-Zhe Guo[1,2†]

[1]*State Key Laboratory of Novel Software Technology, Nanjing University, China*
[2]*School of Intelligence Science and Technology, Nanjing University, China*
[3]*School of Artificial Intelligence, Nanjing University, China*

I. Abstract

While LLM agents demonstrate potential across various domains, their **trustworthiness in complex, real-world tasks** remains under-examined, particularly in travel planning where **constraint satisfaction** acts as a persistent bottleneck. This paper presents a systematic examination of the domain: 1) we provide a **comprehensive review of existing benchmarks**, summarizing design trends and emerging challenges; 2) we categorize prevailing solutions into **general-purpose, multi-agent, and neuro-symbolic approaches**, analyzing their trade-offs; and 3) we introduce **modular ability analyses** to pinpoint failures, revealing that significant challenges remain in **reasoning and processing information under constraints**, suggesting that task decomposition is the most promising path forward.

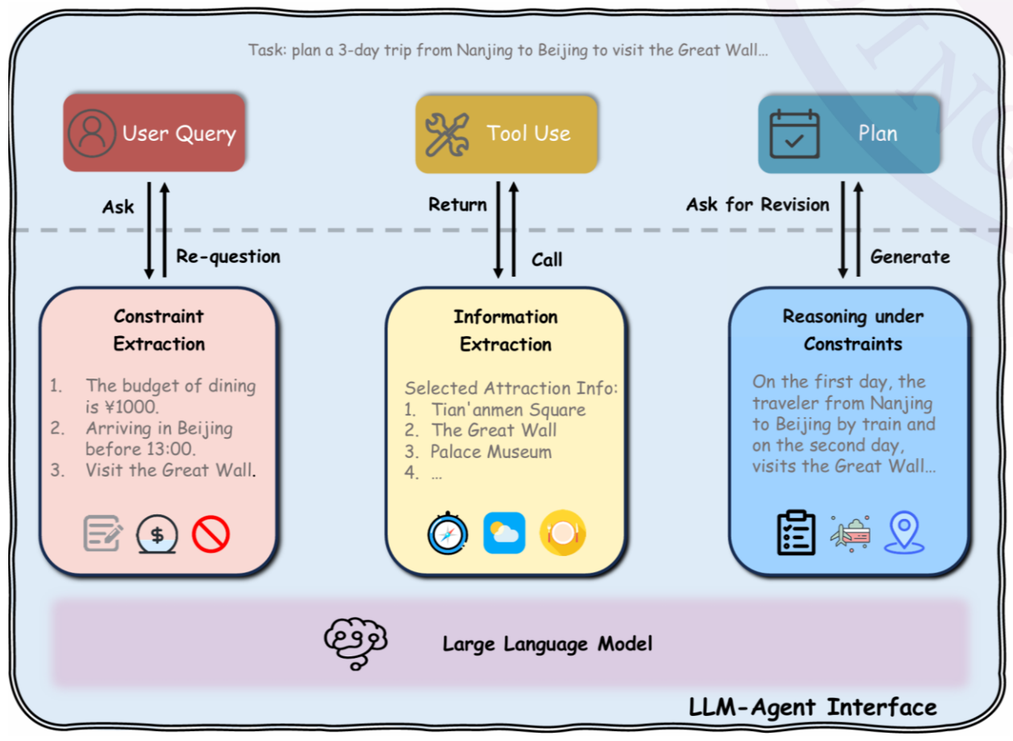
II. Benchmarks Analysis

Benchmark	Goal Interpretation		Information Integration		User-Need Data Design	
	Spatio-Temporal Constraints	Preference Modeling	Static Context	Tool Interection	Human-Authored Queries	Open-World Intent
NATURAL PLAN (Zheng et al. 2024)	✗	✗	✓	✗	✗	✗
TravelPlanner (Xie et al. 2024)	✗	✗	✓	✓	✗	✗
TravelPlanner+ (Singh et al. 2024)	✗	✓	✓	✓	✗	✗
ChinaTravel (Shao et al. 2024)	✓	✓	✗	✓	✓	✓
TripCraft (Chaudhuri et al. 2025)	●	✓	✓	✗	✗	✗
TripTailor (Wang et al. 2025)	✗	✓	✓	✗	✗	✗
RETAIL (Deng et al. 2025)	●	✓	✓	✗	✗	✗
RealTravel (Shao et al. 2025c)	✗	✓	✓	✓	✗	✗
TripTide (Karmakar et al. 2025)	●	✓	✓	✗	✗	✗
TripScore (Qu et al. 2025)	●	✓	✓	✗	✓	✓

✓ Supported ● Partially supported ✗ Not supported

- **Goal Interpretation:** Strict spatio-temporal constraints pose a major challenge to agents' fine-grained reasoning capabilities.
- **Information Integration:** Realistic evaluation benefits from sandbox where agents actively verify dynamic information.
- **User-Need Design:** Implicit constraints in open-world queries remain a critical bottleneck for current models.

III. Key Capabilities



- **Intent Understanding:** Move beyond explicit instruction following to handle **open-world constraints** and **ambiguous user needs** through active clarification and implicit preference extraction.
- **Tool Usage:** Shift from static, closed-world generation to **dynamic** sandbox environments, where agents actively verify real-time information (e.g., availability, pricing) via external APIs.
- **Planning & Reasoning:** As the **central competency**, agents must generate **hallucination-free** itineraries that strictly adhere to **spatio-temporal constraints** while fully satisfying **user requirements**.

IV. Experiments

Overall Results

LLMs struggle to generate entirely valid plans.

Method	Model	DR	Micro-Env	Macro-Env	Micro-Log	Macro-Log	FPR
TravelPlanner-Val							
ReAct	GPT-4-Turbo	89.4	61.1	2.8	15.2	10.6	0.6
DeepResearch	Tongyi	95.0	49.6	0.0	6.19	5.56	0.0
DPPM	DeepSeek-V3	100	96.9	77.8	82.6	73.3	64.4
LLM-Modulo	GPT-4-Turbo	100	89.2	40.6	62.1	39.4	20.6
LLMFP	GPT-4	95.0	95.0	95.0	95.7	98.9	93.3
TTG	GPT-4o	100	85.4	91.7	87.9	91.7	91.7
ChinaTravel-Val							
ReAct	GPT-4o	96.1	50.5	0.0	72.4	32.5	0.0
DeepResearch	Tongyi	24.7	32.5	0.0	57.9	26.6	0.0
LLM-Modulo	GPT-4o	91.5	87.2	3.24	92.9	66.2	3.24
TTG	DeepSeek-V3	9.09	12.8	2.59	7.65	5.19	1.29
TripCraft-Agentic-3-days							
DeepResearch	Tongyi	86.9	55.6	0.0	33.2	21.8	0.0
LLM-Modulo	GPT-4o	100	76.4	0.0	48.0	42.7	0.0
LLMFP	GPT-4o	100	99.3	94.5	99.1	98.8	93.6
TTG	GPT-4o	100	99.0	90.7	94.1	86.6	81.1

Constraints Extraction Results

LLMs are effective at extracting closed world constraints, but open constraint extraction remains challenging.

Benchmark	Model	Precision	Recall	F1 Score
TravelPlanner	GPT-5.2-5.2	0.88	0.87	0.87
	Claude Sonnet 4.5	**0.94**	**0.91**	**0.92**
	Deepseek V3.2	0.91	0.90	0.90
	Qwen3-Max	0.92	0.88	0.90
TripCraft	GPT-5.2-5.2	0.90	0.93	0.92
	Claude Sonnet 4.5	**0.94**	**0.97**	**0.96**
	Deepseek V3.2	0.89	0.92	0.91
	Qwen3-Max	0.92	0.94	0.93
ChinaTravel	GPT-5.2-5.2	0.63	0.58	0.61
	Claude Sonnet 4.5	0.72	**0.65**	**0.69**
	Deepseek V3.2	0.61	0.63	0.62
	Qwen3-Max	**0.74**	0.56	0.64

Error Recognition

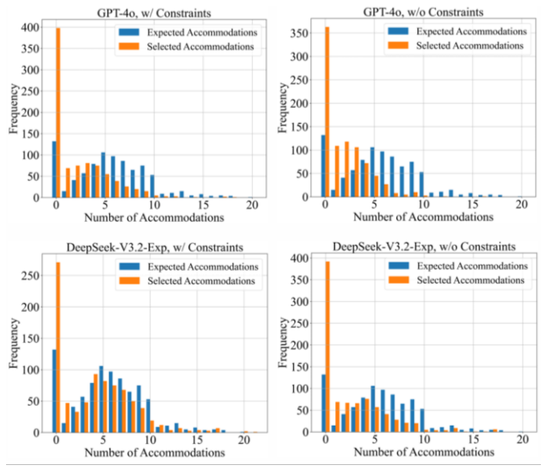
LLMs can recognize some errors given the error range and correct constraints

Model	TravelPlanner	TripCraft	ChinaTravel
DeepSeek-Chat	62.00	**55.00**	66.00
DeepSeek-Reasoner	58.00	54.00	**70.00**
GPT-5.2	44.00	39.00	69.00
GPT-5.2-Medium	**63.00**	43.00	**70.00**

POI Selection

LLMs adopt conservative and inaccurate policies.

Setting	F1	Prec	Rec	EM
GPT-4o				
w/o Const. Ann.	0.27 / 0.36	0.43 / 0.73	0.22 / 0.24	0.09
w/ Const. Ann.	0.32 / 0.43	0.41 / 0.76	0.29 / 0.30	0.18
DeepSeek-V3.2-Exp				
w/o Const. Ann.	0.28 / 0.40	0.34 / 0.65	0.27 / 0.29	0.10
w/ Const. Ann.	0.46 / 0.64	0.50 / 0.76	0.46 / 0.55	0.24



V. Conclusion

- **The Trustworthiness Gap:** A persistent gap remains between current LLM capabilities and robust autonomous planning.
- **The Trade-off:** Existing approaches face a dilemma: Neuro-symbolic methods offer precision but lack adaptability, while General-purpose agents offer broad applicability but fail at complex constraint satisfaction.
- **Future Directions:**
  - **Methodology:** Develop modular frameworks that fuse structured reasoning with adaptive language models.
  - **Evaluation:** Construct realistic, large-scale benchmarks paired with fine-grained diagnostic protocols to accurately isolate specific breakdowns in reasoning and planning.