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
{zbw, yej}@smail.nju.edu.cn, shaojj@lamda.nju.edu.cn, {liyf, guolz}@nju.edu.cn

Abstract

Large language model (LLM) agents are increasingly claimed to handle complex, multi-step tasks, yet their trustworthiness in real-world task remains under-examined. Recent work on travel planning has already pointed out that constraint satisfaction is a persistent bottleneck, especially when itineraries must respect spatio-temporal feasibility, user-specific preferences, and budget or resource limits. However, these observations are mostly made in isolation: they are tied to a single dataset or a particular agent design, which makes it hard to tell whether the weakness is fundamental to current LLM agents or accidental to the setup. This paper presents a systematic examination of travel planning. We present a comprehensive review of existing travel-planning benchmarks, summarizing their design trends and highlighting the new challenges arising from these developments. We also categorize prevailing approaches into general-purpose agent, multi-agent system, and neuro-symbolic approach, and analyze their respective trade-offs between generalizability and domain adaptability. Modular ability analyses are introduced to analyze model performance across them, enabling a deeper investigation into the diverse capabilities required for successful travel planning and revealing the limitations of current methods. We find that significant challenges remain in recognizing open constraints, extracting information under constraints, and reasoning under constraints. Although these complex problems are challenging to tackle as a whole, by decomposing them into manageable sub-tasks, there remains a promising path toward achieving trustworthy agents.

Introduction

In the pursuit of artificial general intelligence (AGI), autonomous agents built upon large language models (LLMs) have emerged as a promising direction (Wang et al. 2024; Yu et al. 2025). In recent years, LLM-based agents equipped with language models capable of perception, reasoning, and decision-making have achieved remarkable progress across diverse domains, including web navigation (Zhou et al. 2024; Deng et al. 2023; Pan et al. 2024), software engineering (Jimenez et al. 2024; Jain et al. 2025), embodied robots (Shridhar et al. 2021; Puig et al. 2018; Srivastava et al. 2022), scientific simulation (Wang et al. 2022; Jansen et al. 2024; Li et al. 2025b). This raises a central question: *are current LLM agents actually trustworthy task executors that can satisfy user-specified goals in real-world settings?*

Recent studies have in fact shown that LLMs still perform poorly in terms of instruction following and meeting user requirements (He et al. 2024; Zhang et al. 2025c; Wen et al. 2024), with travel planning emerging as a particularly revealing domain. In travel planning, given a user query, agents must integrate information from multiple tools (e.g., searching for flights, restaurants, and hotels) to produce a feasible itinerary. Because the reasoning process must jointly account for personal preferences, temporal and spatial dependencies, hard constraints, and real-world factual knowledge, even advanced LLM agents continue to struggle, and their end-to-end success rates can drop close to zero in these realistic settings (Zheng et al. 2024; Xie et al. 2024; Singh et al. 2024; Shao et al. 2024; Valmeekam et al. 2024; Ni et al. 2025; Wang et al. 2025; Chaudhuri et al. 2025; Deng et al. 2025; Qu et al. 2025; Karmakar et al. 2025), which underscores the need for further research to improve reasoning and planning capabilities in complex, real-world scenarios.

In this paper, we revisit the development of the travel planning task. We first review the majority of existing benchmarks, outlining their evolution across three key dimensions, *Goal Interpretation*, *Information Integration* and *User-Need Data Design*, with a detailed discuss the corresponding challenges associated with each. Next, we outline the three main types of solutions currently available. Although general-purpose agents often exhibit suboptimal performance, their broad applicability makes improving their overall capability a worthwhile pursuit. In addition, modular multi-agent systems, which rely on predefined agent workflows, strike a balance between generalization and performance, while neuro-symbolic methods demonstrate strong effectiveness in specific, well-defined (non-open) scenarios. Beyond these, automated agent design and training-based approaches have also been explored. Finally, we conduct relevant experiments and provide fine-grained modular ability analysis. Despite demonstrating strong constraint extraction capability, the model remains limited in its reasoning and information

*These authors contributed equally.
†Corresponding author.

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

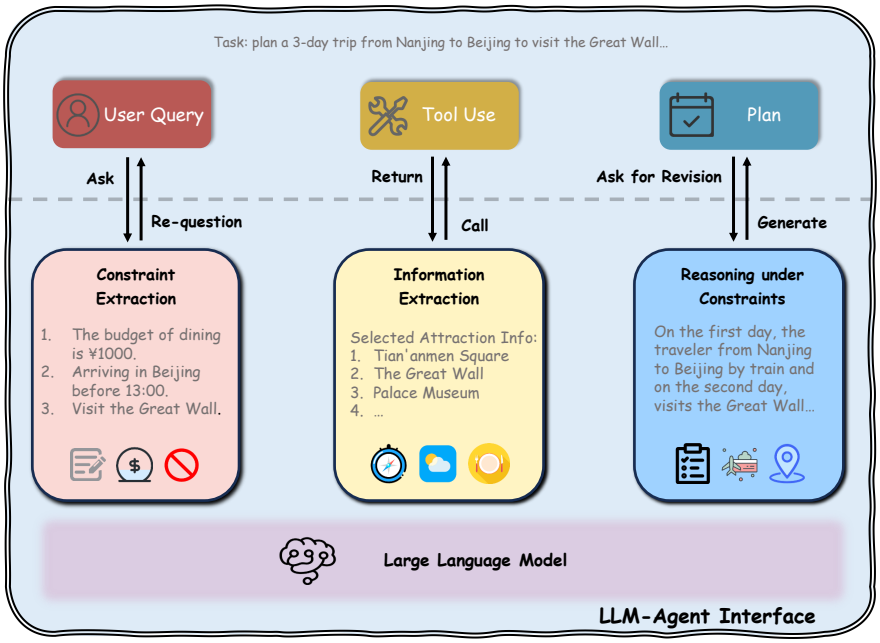

Figure 1: Key Capabilities Required for LLM-Agents to Perform Travel Planning Tasks.

processing when operating under constraints. While there is still no effective universal solution, we hope that by revisiting the performance of these methods on various datasets and analyzing the reasons behind their successes or failures, we can provide valuable insights for future research.

Related Work

LLM-Agents. Empowered by large language models, LLM-Agents can decompose complex tasks and take appropriate actions across diverse scenarios. Representative works include AutoGPT (Yang, Yue, and He 2023), HuggingGPT (Shen et al. 2023), ReAct (Yao et al. 2023), and Reflexion (Shinn et al. 2023). An agent typically consists of several key components: planning, memory, and tool use (Weng 2023). Planning refers to the agent's ability to break down complex tasks and execute them step by step, often adjusting plans in real time based on execution feedback. Existing frameworks have demonstrated strong performance in scenarios. However, despite these successes, their planning strategies remain limited when facing tasks with multiple, interdependent constraints.

Real-World LLM Planning Tasks. Planning is widely recognized as a key hallmark of human intelligence, generally referring to the process of formulating and executing a structured series of actions that lead from an initial to a desired goal state (Hayes-Roth and Hayes-Roth 1979; Grafman, Spector, and Rattermann 2004). Recent research has explored the planning capabilities of LLM-based agents across a wide range of domains (Wei et al. 2025). In benchmarks such as BlocksWorld (Valmeekam et al. 2023), ALFRED (Shridhar et al. 2020), ALFWorld (Shridhar et al. 2021) and VirtualHome (Puig et al. 2018) investigate how agents perceive, reason, and act in interactive environment or simulated physical spaces. In the context of games, environments such as MineCraft, SmartPlay (Wu et al. 2024), and AUCARENA (Chen et al. 2023) provide challenging platforms for testing long-horizon reasoning and adaptive decision-making. For task decomposition, frameworks including TaskLAMA (Yuan et al. 2024) and WORLDAPIS (Ou et al. 2025) examine how LLM agents break down intricate problems into coherent, executable subgoals. Collectively, these studies underscore the rapid progress and continuing challenges in evaluating and enhancing the planning competence of LLM-driven agents across increasingly diverse and realistic settings.

Neuro-Symbolic AI. Neuro-Symbolic Learning seeks to integrate traditional symbolic reasoning with data-driven learning, aiming to enhance both interpretability and reliability (Wang et al. 2019; Manhaeve et al. 2018; Dai et al. 2019; Yang et al. 2024; Shao et al. 2025a; Yang et al. 2025; Shao et al. 2025b). In the era of large language models, (Pan et al. 2023) introduced LogicLM, which combines LLMs with external symbolic solvers to tackle various logical reasoning tasks. In their framework, the LLM first translates a natural language problem into a symbolic representation, after which a deterministic symbolic solver performs inference on the translated problem to ensure correctness. Building on this idea, (Deng, Dong, and Si 2024) augmented LogicLM with a Self-Refinement Module, improving the reliability of the LLM's symbolic translation. In the domain of travel planning, (Hao et al. 2024) proposed a similar neuro-symbolic framework. Their system first extracts logical constraints from natural language queries, then formalizes them into SMT (Satisfiability Modulo Theories) code. Leveraging the soundness and completeness of SMT solvers, this approach guarantees the correctness of gener-

ated plans, achieving an impressive 97% success rate on the TravelPlanner benchmark. Together, these efforts demonstrate how neuro-symbolic AI can complement the generative power of LLMs with formal reasoning capabilities, offering a promising direction for building trustworthy and verifiable intelligent systems.

Benchmarks and Characteristics Analysis

The literature has witnessed an explosive proliferation of travel-planning benchmarks in recent years (Xie et al. 2024; Shao et al. 2024; Tang et al. 2024; Wang et al. 2025; Qu et al. 2025; Esper et al. 2025). From a constraint-satisfaction perspective on whether LLM agents meet user needs, we focus on the widely adopted setting, itinerary generation, and compile representative datasets in the Table 1. Specifically, we analyze these benchmarks through three problem-centric lenses: **Goal Interpretation**, **Information Integration**, and **User-Need Data Design**. We then compare how different benchmarks advance these aspects and discuss consequences for evaluation protocols and model design.

Goal Interpretation

Trustworthiness ultimately hinges on whether an agent can interpret the user's goal and deliver a plan that both satisfies hard constraints and aligns with soft preferences.

Spatiotemporal and Resource Feasibility. On the feasibility side, realistic planning requires spatiotemporal and resource consistency, e.g., calendar-aware scheduling, inter-POI transitions with plausible durations, opening hours, availability, and budgets. TripCraft requires schedules but checks only limited constraints (e.g., safety gaps around inter-city transport), with transportation information that may not reflect actual travel durations; TripTide inherits this design. RETAIL enforces a fixed 30-minute gap between activities, while TripScore adopts point-wise LLM-as-a-Judge to penalize temporal/spatial violations. In contrast, ChinaTravel applies a rigorous rule-based validator to ensure spatiotemporal continuity and non-conflict. Introducing full spatio-temporal constraints also brings additional challenges. First, the spatial dimension requires quadratic-scale transportation data among all POIs, leading to input lengths that often exceed the maximum context window of current LLMs. This is one of the main reasons why ChinaTravel cannot be directly applied to sole-planning tasks. Moreover, from the perspective of a constraint solver, the temporal granularity of variables and constraints shifts from the day level to the hour or even minute level, ideally, the minimal temporal unit, significantly increasing the computational complexity of reasoning.

Preference Modeling. Beyond hard feasibility, realistic planning demands preference modeling and trade-off handling. Benchmarks operationalize soft constraints differently: ChinaTravel formulates preference adherence as optimization over numerical indicators (e.g., total attractions visited); TripCraft measures alignment between soft constraints and POI names via BERT scores; TripScore defines tailored scoring rules per preference type. Several works (e.g., TravelPlanner+, RealTravel, TripTailor) explore LLM-as-a-Judge for preference assessment. These choices materially affect evaluation fidelity and comparability.

Information Integration

Evaluation settings that remove tool invocation implicitly assume a closed, static world in which all travel facts are pre-specified. Real travel planning, however, depends on dynamic information, flight schedules, room availability, prices, opening hours, and service disruptions—that an agent must fetch, verify, and compose to satisfy user needs. Tool use is therefore not an implementation convenience but a core competency of LLM-based agents: accessing external knowledge sources, executing API-level operations, and integrating retrieved signals into reasoning and decision-making (Weng 2023; Liu et al. 2024; Huang et al. 2024). Benchmarks that exclude this dimension risk rewarding in-context text synthesis while obscuring failure modes most relevant to trustworthiness (e.g., availability checks, fact verification, update reconciliation).

To align evaluation with deployment, we adopt a sandbox construction perspective: instrumented, reproducible environments that expose realistic tools (search, booking, maps, events) with observable inputs and outputs, so agents must ground their plans through tool calls. Early work such as TravelPlanner already included a two-stage, tool-using setting, underscoring the role of grounded information integration for realistic planning; ChinaTravel extends this idea with richer and more functionally detailed tools. As our analysis shows, the presence or absence of such sandboxed tools materially changes evaluation protocols and the feasible design space of planning agents.

User-Need Data Design

Early datasets often synthesized "human-like" queries by sampling constraint sets and prompting LLMs to write queries. While this improves formal validity, gaps remain relative to authentic queries, and feasibility is not guaranteed. Subsequent work, e.g., TravelPlanner+ (Singh et al. 2024), TripCraft (Chaudhuri et al. 2025) adopt role-playing/user modeling to improve semantic coherence (e.g., avoiding conflicts between a large budget and consistently low-cost choices). TripTailor (Wang et al. 2025) back-synthesizes queries from structured real itineraries.

Open-World Intent Reasoning. Directly collecting human-authored queries yields the most realistic evaluation but also introduces open-world intent—unbounded, evolving, or implicitly stated requirements beyond a fixed schema (e.g., "local dishes," "kid-friendly," or last-minute changes). This shifts the challenge from template compliance to intent understanding: agents must interpret previously unseen constraints and use them to guide plan generation. In ChinaTravel and TripScore, authentic user requests bring such implicit and diverse expressions, substantially increasing annotation and evaluation difficulty (Shao et al. 2024; Qu et al. 2025). ChinaTravel addresses this with a domain-specific language (DSL) and Python-based annotation to cover arbitrary constraints, while TripScore

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

Table 1: A summary of existing travel-planning benchmarks on constraint satisfaction.

employs LLM-as-a-Judge. The emprical studies from (Shao et al. 2024; Qu et al. 2025) show that open-world intent significantly degrades the performance, and methods relying on explicit constraint extraction can fail catastrophically under such openness. These observations further suggest that evaluation should assess intent resolution (ambiguity detection, targeted clarification, and consistent plan updates) and uncertainty-aware behaviors (calibrated refusals, confidence reporting, and hedged recommendations), rather than only scoring final itineraries.

Miscellaneous

Apart from the above dimensions, several benchmarks explore additional aspects that are not yet systematically studied. TP-RAG (Ni et al. 2025) focuses on the generation of plans under a retrieval-augmented framework, highlighting the impact of external knowledge on planning quality. RETAIL (Deng et al. 2025) investigates the use of simulated user interactions to enhance the modeling of user requirements. TripTide, on the other hand, introduces the task of replanning after interruptions, assessing the agent's capacity to adapt to unforeseen changes.

Methods

In this section, we introduce three distinct approaches to solving the travel planning task: General-purpose agents, Multi-Agent Systems for travel planning, and Neuro-Symbolic approaches. These methods differ in their design philosophies, reliance on human knowledge, and the complexity of problem-solving they entail. Table 2 presents a selection of travel planning methods and their corresponding categories. Other approaches, such as training-based methods and automated agent design, will be discussed at the end of this section.

General-Purpose Agent

General-Purpose agents are not tailored to specific tasks or domains. Instead, they emphasize versatility and generalization across a wide range of scenarios.

In the two-stage setting, agents such as ReAct (Yao et al. 2023) and Reflexion (Shinn et al. 2023) aim to minimize restrictions on the LLM's behavior, leveraging its intrinsic planning and reasoning capabilities to complete tasks. Reflexion further utilizes the intrinsic capabilities of LLMs to summarize errors and refine its experiences for continuous improvement. While this method enables steady performance improvements with repeated attempts across various tasks, it fails to yield significant improvements in travel planning and may even cause performance degradation.

These methods generally have low reliance on human knowledge. They typically require only basic task description, enabling rapid adaptation and transfer across various tasks and domains.

Multi-Agent System

Multi-agent systems in travel planning are typically designed with multiple task-specific agents, which decompose complex tasks into simpler subtasks to ultimately generate a valid itinerary.

HLRF implements a human-like planning framework, decomposing the task into three stages: Outline Generation, Information Collection, and Plan Making, with multiple agents collaborating within each stage to complete the corresponding subtasks. ISP introduces an approximation agent to realize a hierarchical planning approach. PMC adopts a relatively flexible framework, consisting of four types of agents: Manager, Executors, Supervisor, and Deliverer. In DPPM, different types of planning information are handled by specialized agents and later merged into a final plan. TGMA also employs a hierarchical planning strategy, where multiple agents iteratively refine the plan.

Compared to general-purpose agents, these multi-agent systems achieve absolute improvements in success rates on travel planning tasks ranging from roughly 2–3% up to over 80%. The extent of improvement typically depends on the degree to which human prior knowledge is incorporated into the multi-agent design and the specific difficulty of the benchmark. However, such methods are often designed for specific benchmarks and tend to exhibit limited transferability across different benchmarks.

Category	Methods
General-Purpose Agent	ReAct (Yao et al. 2023), Reflexion (Shinn et al. 2023) Tongyi-DeepResearch (Li et al. 2025a)
Multi-Agent System	HLRF (Xie and Zou 2024), ISP (Hua et al. 2025), PMC (Zhang et al. 2025a), DPPM (Lu et al. 2025), TGMA (Deng et al. 2025)
Neuro-Symbolic	LLM-Modulo (Kambhampati et al. 2024), PTS (Shao et al. 2025c), ChinaTravel-NeSy (Shao et al. 2024), LLMFP (Hao et al. 2024), TTG (Ju et al. 2024)

Table 2: Representative methods for travel planning tasks.

Neuro-Symbolic Approach

Neuro-symbolic approaches combine LLMs with symbolic systems, offering a potential pathway toward robust and trustworthy planning methods. They typically achieve better performance compared to the two aforementioned types of methods.

LLM-Modulo is a relatively general approach that requires a verifier capable of checking answers for a given task and providing error feedback, with the LLM iteratively refining the plan based on this feedback. PTS consists of five sequential modules, Translation, Search, Preference, Re-rank, and Planning, which work together to produce high-quality final plans. The NeSy method in ChinaTravel is an LLM-guided search approach designed to reduce the search space. LLMFP and TTG are both LLM+Solver approaches, modeling problems as SMT and MILP instances, respectively, and leveraging existing solvers for solution. It is important to note that these methods heavily rely on expert-crafted prompts, which essentially amount to encoding the logic of an expert system for problem solving. Nonetheless, they achieve excellent efficiency and success rates during execution. To prevent the model from generating incorrect code, solver-based approaches often require carefully designed pipelines that provide step-by-step prompts and merge the resulting code at the end.

Despite their strong performance, the generalizability of neuro-symbolic methods remains a concern. Those LLM+Solver approaches that emphasize zero-shot generalization are evaluated on other relatively simpler tasks (Hao, Zhang, and Fan 2025). Their robustness is limited when facing challenges such as open constraints (Qu et al. 2025), and studies on ChinaTravel further indicate that their performance degrades significantly as task complexity increases.

Miscellaneous

Additionally, some automatically designed agents can be used in these tasks as well (Shang et al. 2025). Some approaches have also modified the task setup of TravelPlanner, using training-based methods to improve performance (Zhang et al. 2025b).

Experiments and Results

Experimental Setup

Due to resource constraints and the complexity of the task, we selected subsets from three benchmarks as our experimental datasets. The methods we explored include direct prompting of the model, LLM-modulo, ReAct, LLMFP, and TTG. Moreover We adapted Tongyi's DeepResearch, utilizing its Tongyi-DeepResearch-30B-A3B model, aiming to leverage the DeepResearch capabilities of the model for generating the final plans (Li et al. 2025a). For the results already available in the original benchmark or method papers, we directly reported the existing results. For TripCraft, which does not fully support tool usage, we implemented a full tool setup modeled after TravelPlanner.

Metrics

In all three benchmarks, we report: (1) Delivery Rate, the rate at which valid responses are generated; (2) Micro/Macro Commonsense Pass Rates, the extent to which outputs adhere to commonsense; (3) Micro/Macro Hard Constraint Pass Rates, the extent to which plans satisfy task-specific, user-defined, or domain-mandated requirements; and (4) Final Pass Rate, the proportion of responses satisfying all criteria simultaneously.

Main Results

The main experimental results are presented in Table 3.

> **Finding 1.** In small-scale tasks, neuro-symbolic approaches outperform purely neural agent methods. However, in large-scale tasks, even LLM+Solver methods suffer from significant performance degradation.

The experimental results reveal a pronounced performance disparity across benchmarks. On TravelPlanner and TripCraft-3days, neuro-symbolic approaches—such as LLM-Modulo, TTG, and LLMFP—consistently outperform purely neural agentic methods by a significant margin. This divergence underscores a fundamental limitation of large language models (LLMs) in handling structured, domain-specific reasoning tasks, which can be effectively mitigated through integration with formal symbolic reasoning mechanisms. In contrast, on the ChinaTravel benchmark, these neuro-symbolic systems exhibit a marked decline in performance. This indicates that in large-scale tasks, neuro-symbolic methods also face considerable challenges.

Notably, despite achieving high scores on current evaluation metrics, LLM+Solver methods often produce travel plans that are technically valid yet practically implausible or unrealistic (e.g., exhibiting poor spatial and temporal coherence or being too strict in format). This discrepancy sug-

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

Table 3: Comparison of Methods across Different Benchmarks.

gests that existing benchmarks without spatial and temporal constraints are insufficiently equipped to assess the real-world feasibility, contextual appropriateness, and overall coherence of generated itineraries.

Moreover, prevailing evaluation protocols tend to assess planning capability through coarse-grained, aggregate metrics, such as adherence to commonsense norms and satisfaction of hard constraints, thereby conflating distinct cognitive and operational competencies essential to effective travel planning. These include, but are not limited to: precise tool invocation, multi-step causal and temporal reasoning, dynamic coordination across activities, and adaptive replanning in response to perturbations. To enable more diagnostic and meaningful model comparisons, future benchmarks should adopt a fine-grained evaluation framework that explicitly disentangles and independently measures these constituent planning skills. Such a framework would provide a more accurate and interpretable assessment of a system's true planning proficiency.

Fine-Grained Modular Ability Analysis

In travel planning, we conduct a comprehensive evaluation of the model's capabilities in constraint understanding, long-horizon reasoning, tool use, and related skills. However, agentic approaches perform unexpectedly poorly on this task, making it difficult to determine which specific challenges limit the language model's performance. We can further decompose the travel planning task into a set of ability modules to investigate why agents struggle to handle such tasks effectively. We investigate the following two tasks:

Constraint Extraction

Constraint extraction requires models to identify and output the corresponding constraints purely from natural language. This is considered a key component of the travel planning task. Only when a model can accurately extract such constraints can it subsequently generate plans that adhere to them.

> **Finding 2**. Existing SOTA LLMs have little problem extracting predefined constraints, while it struggles with diverse expressions and open-ended constraints.

Experimental results show that language models can almost perfectly extract these predefined constraints on TravelPlanner: GPT-4o and DeepSeek-V3.2-Exp achieve accuracies of 99.89% and 99.78%, respectively, which essentially indicates that the models are fully capable of extracting all predefined constraints, with only a few ambiguous cases leading to mismatches with the annotations.

Both models failed on a query containing the confusing expression "We require accommodations that are neither shared nor subject to visitor restrictions and should be private rooms.", which redundantly specifies both "not shared room" and "private room." Additionally, DeepSeek-V3.2-Exp produced a result differing from the annotation for the query "should ideally be non-shared rooms," where the annotated constraint was "private room."

Notably, ChinaTravel conducted a similar analysis on its POI Reasoning task, which can be seen as a simplified form of constraint extraction. In this task, the model is required to fill in the corresponding values for constraints where specific values have been removed. The results indicate that extracting constraints from ChinaTravel's open-ended queries is extremely challenging.

Constraint-based Information Extraction

According to the constraint-based information extraction setting, the model is required to extract useful information from the tool's returned results, reflecting its ability to reason under constraints. We designed a simple experiment

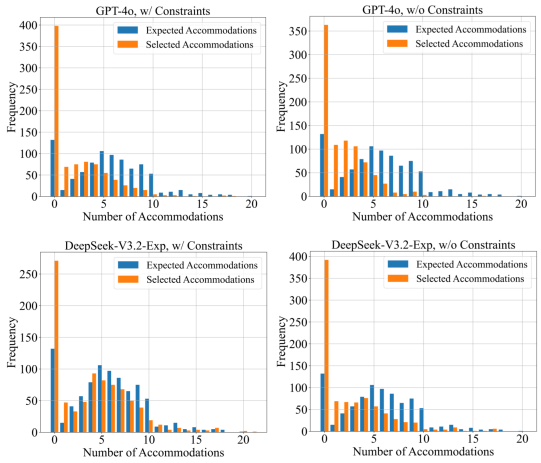

Figure 2: Oracle vs LLM: Distribution of Accommodations Satisfying the Constraints.

where the model was asked to select feasible accommodations only based on the room type and house rule constraints, without considering budget or minimum stay duration. Specifically, for each query that includes a room type or house rule, we retrieve accommodations from all possible cities related to the query's destination, and ask the model to select those that meet both requirements. In real-world tasks, however, each selected POI must not only satisfy its own constraints but also maintain consistency with others, making the reasoning process much more complex.

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

Table 4: Results on *Constraint-based Info. Extraction (Accommodation Selection)*. EM = Exact Match Ratio. Macro / Micro metrics are shown as 'Macro / Micro'.

Finding 3. Even with ground-truth constraints, SOTA LLMs still struggle to apply them during the reasoning.

As shown in Table 4, neither model was able to select all candidates that fully met the given requirements, indicating that in the full task, potentially correct plans may be prematurely discarded. Conversely, both models also failed to exclude all unsuitable options, suggesting that erroneous items might be incorporated into the final plan. LLMs showed a clear improvement when constraint annotations were available, implying that even with a strong capability for constraint extraction, they still tend to overlook constraint information expressed in natural language when explicit annotations are absent during reasoning.

Finding 4. The SOTA LLMs tend to adopt a conservative strategy, yet they still struggle to fully satisfy all constraints, often leading to false positives despite its cautious approach.

Based on Figure 2, we can make an interesting observation: the model tends to adopt a conservative strategy, typically selecting fewer hotels. This tendency is also reflected in the noticeably higher precision compared to recall. However, the large gap between the macro and micro versions of precision further indicates that such a conservative strategy does not necessarily lead to better accuracy; instead, it results in a higher proportion of false positives among the smaller set of candidate hotels.

Reasoning under Constraints

Finding 5. Even for Large Reasoning Models (LRMs), reasoning under constraints remains challenging.

Reasoning under constraints requires models to directly infer the final plan given the available information. We collected results from LLM-Modulo and TravelPlanner, and observed that while performance slightly improves as model reasoning ability increases, it remains relatively low overall, indicating that there is still substantial room for improvement in models' reasoning capabilities under constraints.

Model	LRM	Pass Rate
GPT-3.5-Turbo	✗	0
GPT-4-Turbo	✗	4.40
o1-mini	✓	1.67
o1-preview	✓	10.0

Table 5: Final pass rate on TravelPlanner in direct setting.

Conclusion

Our findings highlight a persistent gap between current LLM-based agents and trustworthy autonomous planning systems. While neuro-symbolic hybrids demonstrate clear advantages in structured reasoning and constraint satisfaction, their reliance on human-engineered priors limits adaptability. Conversely, general-purpose agents exhibit broader transferability but lack precision in handling complex, interdependent constraints. The divergence across benchmarks suggests that progress toward reliable travel planning agents requires both methodological innovation and evaluative reform. Promising future directions include (1) developing modular frameworks that combine structured reasoning with adaptive language models, (2) constructing large-scale, realistic benchmarks that better reflect user interaction and spatiotemporal complexity, and (3) designing fine-grained and interpretable evaluation protocols to diagnose specific reasoning and planning capabilities. Collectively, these directions may foster more reliable, transparent, and trustworthy LLM-based agents capable of robust real-world planning.

References

Chaudhuri, S.; Purkar, P.; Raghav, R.; Mallick, S.; Gupta, M.; Jana, A.; and Ghosh, S. 2025. Tripcraft: A benchmark for spatio-temporally fine grained travel planning. *arXiv preprint arXiv:2502.20508*.

Chen, J.; Yuan, S.; Ye, R.; Majumder, B. P.; and Richardson, K. 2023. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

Dai, W.-Z.; Xu, Q.; Yu, Y.; and Zhou, Z.-H. 2019. Bridging machine learning and logical reasoning by abductive learning. *Advances in Neural Information Processing Systems*.

Deng, B.; Feng, Y.; Liu, Z.; Wei, Q.; Zhu, X.; Chen, S.; Guo, Y.; and Wang, Y. 2025. Retail: Towards real-world travel planning for large language models. *arXiv preprint arXiv:2508.15335*.

Deng, S.; Dong, H.; and Si, X. 2024. Enhancing and evaluating logical reasoning abilities of large language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*.

Esper, J. P.; Fraga, L. d. S.; Viana, A. C.; Cardoso, K. V.; and Correa, S. L. 2025. + Tour: Recommending personalized itineraries for smart tourism. *Computer Networks*, 260: 111118.

Grafman, J.; Spector, L.; and Rattermann, M. J. 2004. Planning and the brain. In *The Cognitive Psychology of Planning*, 191–208. Psychology Press.

Hao, Y.; Chen, Y.; Zhang, Y.; and Fan, C. 2024. Large Language Models Can Solve Real-World Planning Rigorously with Formal Verification Tools. *arXiv preprint arXiv:2404.11891*.

Hao, Y.; Zhang, Y.; and Fan, C. 2025. Planning Anything with Rigor: General-Purpose Zero-Shot Planning with LLM-based Formalized Programming. In *Proceedings of the 13th International Conference on Learning Representations*.

Hayes-Roth, B.; and Hayes-Roth, F. 1979. A cognitive model of planning. *Cognitive Science*, 3(4): 275–310.

He, Q.; Zeng, J.; Huang, W.; Chen, L.; Xiao, J.; He, Q.; Zhou, X.; Liang, J.; and Xiao, Y. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 18188–18196.

Hua, W.; Wan, M.; VADREVU, J. S. S. S.; Nadel, R.; Zhang, Y.; and Wang, C. 2025. Interactive Speculative Planning: Enhance Agent Efficiency through Co-design of System and User Interface. In *Proceedings of the 13th International Conference on Learning Representations*.

Huang, X.; Liu, W.; Chen, X.; Wang, X.; Wang, H.; Lian, D.; Wang, Y.; Tang, R.; and Chen, E. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716*.

Jain, N.; Han, K.; Gu, A.; Li, W.-D.; Yan, F.; Zhang, T.; Wang, S.; Solar-Lezama, A.; Sen, K.; and Stoica, I. 2025. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *Proceedings of the 13th International Conference on Learning Representations*.

Jansen, P.; Côté, M.-A.; Khot, T.; Bransom, E.; Dalvi Mishra, B.; Majumder, B. P.; Tafjord, O.; and Clark, P. 2024. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. R. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *Proceedings of the 12th International Conference on Learning Representations*.

Ju, D.; Jiang, S.; Cohen, A.; Foss, A.; Mitts, S.; Zharmagambetov, A.; Amos, B.; Li, X.; Kao, J. T.; Fazel-Zarandi, M.; and Tian, Y. 2024. 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*, 240–249.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L. P.; and Murthy, A. B. 2024. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, 22895–22907.

Karmakar, P.; Chaudhuri, S.; Mallick, S.; Gupta, M.; Jana, A.; and Ghosh, S. 2025. TripTide: A Benchmark for Adaptive Travel Planning under Disruptions. *arXiv preprint arXiv:2510.21329*.

Li, B.; Zhang, B.; Zhang, D.; Huang, F.; Li, G.; Chen, G.; Yin, H.; Wu, J.; Zhou, J.; et al. 2025a. Tongyi DeepResearch Technical Report. *arXiv preprint arXiv:2510.24701*.

Li, R.; Hu, Z.; Qu, W.; Zhang, J.; Yin, Z.; Zhang, S.; Huang, X.; Wang, H.; Wang, T.; Pang, J.; Ouyang, W.; Bai, L.; Zuo, W.; Duan, L.-Y.; Zhou, D.; and Tang, S. 2025b. LabUtopia: High-Fidelity Simulation and Hierarchical Benchmark for Scientific Embodied Agents. *arXiv preprint arXiv:2505.22634*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2024. AgentBench: Evaluating LLMs as Agents. In *Proceedings of the 12th International Conference on Learning Representations*.

Lu, Z.; Lu, W.; Tao, Y.; Dai, Y.; Chen, Z.; Zhuang, H.; Chen, C.; Peng, H.; and Zeng, Z. 2025. Decompose, Plan in Parallel, and Merge: A Novel Paradigm for Large Language Models based Planning with Multiple Constraints. *arXiv preprint arXiv:2506.02683*.

Manhaeve, R.; Dumancic, S.; Kimmig, A.; Demeester, T.; and De Raedt, L. 2018. Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*.

Ni, H.; Liu, F.; Ma, X.; Su, L.; Wang, S.; Yin, D.; Xiong, H.; and Liu, H. 2025. TP-RAG: Benchmarking Retrieval-Augmented Large Language Model Agents

for Spatiotemporal-Aware Travel Planning. *arXiv preprint arXiv:2504.08694*.

Ou, J.; Uzunoğlu, A.; Van Durme, B.; and Khashabi, D. 2025. Worldapis: The world is worth how many apis? a thought experiment. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 24993–25001.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3806–3824.

Pan, Y.; Kong, D.; Zhou, S.; Cui, C.; Leng, Y.; Jiang, B.; Liu, H.; Shang, Y.; Zhou, S.; Wu, T.; and Wu, Z. 2024. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*.

Puig, X.; Ra, K.; Boben, M.; Li, J.; Wang, T.; Fidler, S.; and Torralba, A. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8494–8502.

Qu, Y.; Xiao, H.; Li, F.; Zhou, H.; and Dai, X. 2025. TripScore: Benchmarking and rewarding real-world travel planning with fine-grained evaluation. *arXiv preprint arXiv:2510.09011*.

Shang, Y.; Li, Y.; Zhao, K.; Ma, L.; Liu, J.; Xu, F.; and Li, Y. 2025. AgentSquare: Automatic LLM Agent Search in Modular Design Space. In *Proceedings of the 13th International Conference on Learning Representations*.

Shao, J.-J.; Hao, H.-R.; Yang, X.-W.; and Li, Y.-F. 2025a. Abductive Learning for Neuro-Symbolic Grounded Imitation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1221–1232.

Shao, J.-J.; Yang, X.-W.; Zhang, B.-W.; Chen, B.; Wei, W.-D.; Guo, L.-Z.; and Li, Y.-f. 2024. ChinaTravel: A Real-World Benchmark for Language Agents in Chinese Travel Planning. *arXiv preprint arXiv:2412.13682*.

Shao, J.-J.; You, H.-J.; Cai, G.; Dai, Q.; Dong, Z.; and Guo, L.-Z. 2025b. Breaking the Self-Evaluation Barrier: Reinforced Neuro-Symbolic Planning with Large Language Models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 6129–6137.

Shao, Z.; Wu, J.; Chen, W.; and Wang, X. 2025c. Personal Travel Solver: A Preference-Driven LLM-Solver System for Travel Planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 27622–27642.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10740–10749.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the 9th International Conference on Learning Representations*.

Singh, H.; Verma, N.; Wang, Y.; Bharadwaj, M.; Fashandi, H.; Ferreira, K.; and Lee, C. 2024. Personal large language model agents: A case study on tailored travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 486–514.

Srivastava, S.; Li, C.; Lingelbach, M.; Martín-Martín, R.; Xia, F.; Vainio, K. E.; Lian, Z.; Gokmen, C.; Buch, S.; Liu, K.; Savarese, S.; Gweon, H.; Wu, J.; and Li, F.-F. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, 477–490.

Tang, Y.; Wang, Z.; Qu, A.; Yan, Y.; Wu, Z.; Zhuang, D.; Kai, J.; Hou, K.; Guo, X.; Zhao, J.; Zhao, Z.; and Ma, W. 2024. ItiNera: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1413–1432.

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*.

Valmeekam, K.; Stechly, K.; Gundawar, A.; and Kambhampati, S. 2024. Planning in strawberry fields: Evaluating and improving the planning and scheduling capabilities of lrm o1. *arXiv preprint arXiv:2410.02162*.

Wang, K.; Shen, Y.; Lv, C.; Zheng, X.; and Huang, X.-J. 2025. Triptailor: A real-world benchmark for personalized travel planning. In *Findings of the Association for Computational Linguistics*, 9705–9723.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wang, P.-W.; Donti, P.; Wilder, B.; and Kolter, Z. 2019. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proceedings of the 36th International Conference on Machine Learning*, 6545–6554.

Wang, R.; Jansen, P.; Côté, M.-A.; and Ammanabrolu, P. 2022. ScienceWorld: Is your Agent Smarter than a 5th Grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11279–11298.

Wei, H.; Zhang, Z.; He, S.; Xia, T.; Pan, S.; and Liu, F. 2025. Plangenllms: A modern survey of llm planning capabilities. *arXiv preprint arXiv:2502.11221*.

Wen, B.; Ke, P.; Gu, X.; Wu, L.; Huang, H.; Zhou, J.; Li, W.; Hu, B.; Gao, W.; Xu, J.; Liu, Y.; Tang, J.; Wang, H.; and Huang, M. 2024. Benchmarking complex instruction-following with multiple constraints composition. *Advances in Neural Information Processing Systems*.

Weng, L. 2023. LLM-powered Autonomous Agents. *lilianweng.github.io*.

Wu, Y.; Tang, X.; Mitchell, T.; and Li, Y. 2024. SmartPlay: A Benchmark for LLMs as Intelligent Agents. In *Proceedings of the 12th International Conference on Learning Representations*.

Xie, C.; and Zou, D. 2024. A human-like reasoning framework for multi-phases planning task with large language models. *arXiv preprint arXiv:2405.18208*.

Xie, J.; Zhang, K.; Chen, J.; Zhu, T.; Lou, R.; Tian, Y.; Xiao, Y.; and Su, Y. 2024. TravelPlanner: a benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning*, 54590–54613.

Yang, H.; Yue, S.; and He, Y. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Yang, X.-W.; Shao, J.-J.; Guo, L.-Z.; Zhang, B.-W.; Zhou, Z.; Jia, L.-H.; Dai, W.-Z.; and Li, Y.-F. 2025. Neuro-symbolic artificial intelligence: towards improving the reasoning abilities of large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 10770–10778.

Yang, X.-W.; Wei, W.-D.; Shao, J.-J.; Li, Y.-F.; and Zhou, Z.-H. 2024. Analysis for Abductive Learning and Neural-Symbolic Reasoning Shortcuts. In *Proceedings of the 41st International Conference on Machine Learning*, 56524–56541.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*.

Yu, M.; Meng, F.; Zhou, X.; Wang, S.; Mao, J.; Pan, L.; Chen, T.; Wang, K.; Li, X.; Zhang, Y.; An, B.; and Wen, Q. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6216–6226.

Yuan, Q.; Kazemi, M.; Xu, X.; Noble, I.; Imbrasaite, V.; and Ramachandran, D. 2024. Tasklama: probing the complex task understanding of language models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 19468–19476.

Zhang, C.; Goh, X. D.; Li, D.; Zhang, H.; and Liu, Y. 2025a. Planning with multi-constraints via collaborative language agents. In *Proceedings of the 31st International Conference on Computational Linguistics*, 10054–10082.

Zhang, K.; Chen, X.; Liu, B.; Xue, T.; Liao, Z.; Liu, Z.; Wang, X.; Ning, Y.; Chen, Z.; Fu, X.; et al. 2025b. Agent Learning via Early Experience. *arXiv preprint arXiv:2510.08558*.

Zhang, T.; Zhu, C.; Shen, Y.; Luo, W.; Zhang, Y.; Liang, H.; Yang, F.; Lin, M.; Qiao, Y.; Chen, W.; Cui, B.; Zhang, W.; and Zhou, Z. 2025c. Cfbench: A comprehensive constraints-following benchmark for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 32926–32944.

Zheng, H. S.; Mishra, S.; Zhang, H.; Chen, X.; Chen, M.; Nova, A.; Hou, L.; Cheng, H.-T.; Le, Q. V.; Chi, E. H.; and Zhou, D. 2024. Natural plan: Benchmarking llms on natural language planning. *arXiv preprint arXiv:2406.04520*.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *Proceedings of the 12th International Conference on Learning Representations*.