

Demystifying Reinforcement Learning for Long-Horizon Tool-Using Agents: A Comprehensive Recipe

Xixi Wu¹ Qianguo Sun² Ruiyang Zhang³ Chao Song²

Junlong Wu² Yiyan Qi²(✉) Hong Cheng¹(✉)

¹The Chinese University of Hong Kong ²IDEA Research ³University of Macau

{xxwu, hcheng}@se.cuhk.edu.hk qiyiyan@idea.edu.cn

 Code  Models

Abstract

Reinforcement Learning (RL) is essential for evolving Large Language Models (LLMs) into autonomous agents capable of long-horizon planning, yet a practical recipe for scaling RL in complex, multi-turn environments remains elusive. This paper presents a systematic empirical study using TravelPlanner, a challenging testbed requiring tool orchestration to satisfy multifaceted constraints. We decompose the agentic RL design space along 5 axes: reward shaping, model scaling, data composition, algorithm selection, and environmental stability. Our controlled experiments yield 7 key take-aways, e.g., (1) reward and algorithm choices are scale-dependent as smaller models benefit from staged rewards and enhanced exploration, whereas larger models converge efficiently with simpler dense rewards, (2) ~ 1 K training samples with a balanced difficulty mixture mark an empirical sweet spot for balancing in-domain and out-of-domain performance under our primary setting, and (3) environmental stability is critical to prevent policy degradation. Based on our distilled recipe, our RL-trained models achieve state-of-the-art performance on TravelPlanner, outperforming leading LLMs.

1 Introduction

Large Language Models (LLMs) have evolved from static text generators into general-purpose autonomous agents capable of reasoning, acting, and interacting with dynamic environments (Yao et al., 2023; Schick et al., 2023). This paradigm shift has enabled diverse applications, ranging from information-seeking agents navigating open-ended web environments (Jin et al., 2025b; Li et al., 2025) to GUI agents manipulating complex user interfaces (Xie et al., 2024b) and software engineering agents modifying and debugging real-world codebases (Jimenez et al., 2024; Yang et al., 2025a, 2024). Across these scenarios, agents must engage in long-horizon planning: decomposing high-level

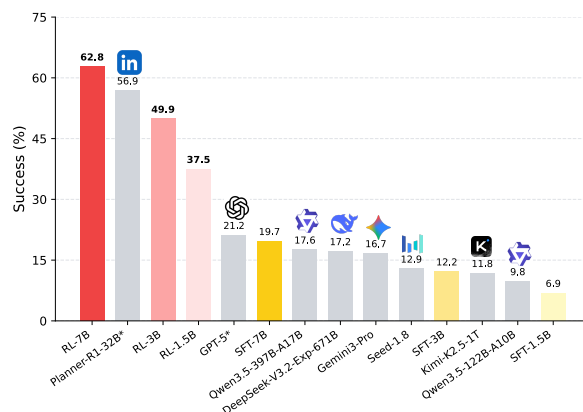


Figure 1: **Performance on the TravelPlanner test set.** Our **RL-trained models** significantly outperform both their **SFT counterparts** and **leading LLMs**. Results marked with * are sourced from Zhu et al. (2025).

goals into manageable sub-tasks, orchestrating tool usage, and satisfying multifaceted constraints to ensure the successful completion of tasks (Xie et al., 2024a; Wu et al., 2024).

Training agents capable of long-horizon tool use remains an open challenge, establishing Reinforcement Learning (RL) as a primary paradigm for optimizing these capabilities through exploration and feedback (AI, 2025; Li et al., 2026a). However, existing insights into agentic RL stem predominantly from short-horizon tasks involving single-step reasoning (Yang et al., 2025b) or few-turn interactions (Jin et al., 2025a; Yu et al., 2025b). In contrast, real-world agentic workflows require long-horizon planning, characterized by dozens of tool invocations and extensive trajectories. While recent efforts have introduced targeted algorithms to tackle this complexity, such as modifying exploration strategies (Ji et al., 2026; Dong et al., 2026) or synthesizing adaptive environments (Lu et al., 2026; Wang et al., 2026b), these works typically explore a limited subset of the RL design space. Crucially, they lack a holistic view of how factors ranging from reward shaping and data composition to model scaling and environmental stability jointly

shape performance. Therefore, the community still lacks a systematic and practical recipe for scaling RL in complex, long-horizon agentic scenarios.

To fully explore this design space and bridge the aforementioned gap, we require an environment that is both complex yet computationally tractable. We adopt **TravelPlanner** (Xie et al., 2024a) as our primary testbed, which captures several core challenges of long-horizon agents. It requires **orchestrating diverse tools** (e.g., transport and accommodation search) to satisfy **multifaceted constraints** (e.g., budget, personal preferences, and hallucination avoidance), presenting a challenge where even top-tier models such as Kimi-K2.5 (Team, 2026a) achieve success rates below 15%. Unlike tasks relying on costly and high-latency external APIs, TravelPlanner operates within a local sandbox, providing a controlled, reproducible, zero-cost, and high-throughput environment essential for scaling RL exploration. Leveraging this efficient testbed, we implement **STAR** (Synthesis, Training, And Reinforcement), a unified post-training pipeline that serves as a scaffold for controlled comparisons across RL design choices. Furthermore, moving beyond intra-task evaluation, we assess our trained policies on both in-domain planning tasks and out-of-domain (OOD) knowledge-intensive QA and deep-research-style information-seeking benchmarks (Chen et al., 2025) to evaluate their broader generalization.

Utilizing the STAR framework, we conduct a large-scale empirical study to decompose the long-horizon RL design space along 5 critical axes: **reward shaping** (dense vs. sparse, with or without curriculum-style staging), **model scaling** (1.5B, 3B, and 7B variants), **data composition** (sample quantity and difficulty), **algorithm selection** (standard GRPO vs. exploration-heavy variants), and **environmental stability** (injecting tool failures and observation corruption). Across these axes, we train **over 30 distinct RL agents**, totaling over **6,500 A100 GPU hours**. By rigorously isolating each factor, we distill the following key takeaways: (1) **Reward and algorithm choices are scale-dependent**: smaller models benefit most from staged curriculum rewards and exploration-heavy algorithms, whereas larger models favor simpler dense rewards and standard GRPO for both accuracy and efficiency; (2) **Data exhibits a sweet spot**: in our primary setting, approximately 1K training samples with a balanced difficulty mixture provide a strong trade-off between in-domain

performance and OOD generalization; and (3) **Environmental stability is critical**: environmental noise can noticeably degrade the performance of long-horizon agents. Finally, following the optimal strategies identified across these factors, our STAR-trained 1.5B-7B models achieve state-of-the-art (SOTA) performance on the TravelPlanner test set, significantly outperforming the strongest commercial LLMs as shown in Figure 1.

In summary, our contributions are as follows:

- **A Unified Post-training Pipeline**: We leverage TravelPlanner as a scalable testbed for long-horizon agents and instantiate STAR, a unified pipeline encompassing data synthesis, supervised fine-tuning (SFT), and RL, providing a scaffold for systematic empirical analysis of agentic RL.
- **A Large-scale Empirical Study**: We systematically dissect the RL design space, providing empirical insights into how reward shaping, model scaling, data composition, algorithm selection, and environmental stability shape policy optimization.
- **Actionable Recipe & SOTA Performance**: We derive a practical, scale-aware recipe for training long-horizon agents. Applying this recipe, our open-weight models achieve SOTA performance on TravelPlanner, surpassing leading proprietary LLMs and providing a foundation for future agentic RL research.

2 Preliminaries

We use **TravelPlanner** (Xie et al., 2024a) as our primary testbed. This platform simulates a realistic travel agency scenario where agents execute long-horizon planning under multifaceted constraints. These constraints encompass both *explicit user requirements* (e.g., budget limits and personal preferences) and *implicit commonsense rules* (e.g., factual grounding and logical consistency). The testbed provides 6 information-gathering tools (e.g., SearchFlight) that query a large-scale local database, with detailed statistics provided in **Appendix B** and Table 6. This configuration preserves realistic multi-tool planning complexity while ensuring the zero-cost, low-latency interactions essential for scalable RL.

ReAct Inference: As illustrated in Figure 2, we employ the ReAct paradigm (Yao et al., 2023) to

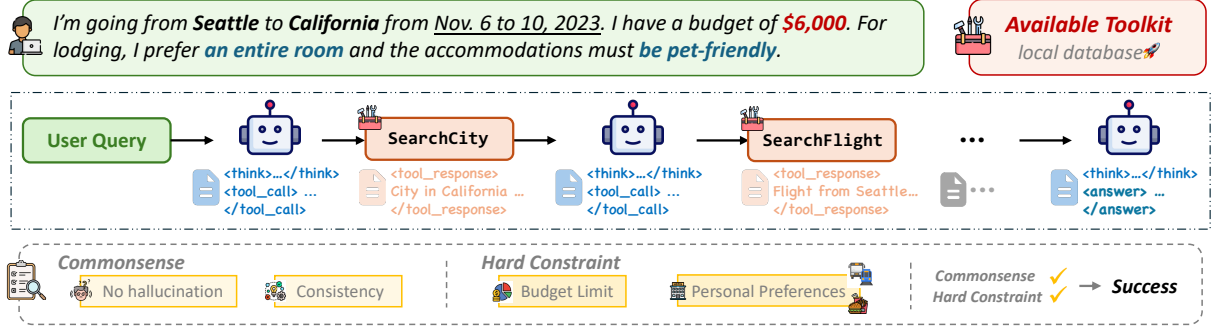


Figure 2: **Overview of the TravelPlanner testbed.** **Top:** A user query alongside the available local toolkit. **Middle:** The agent’s iterative reasoning and tool-use trajectory following the ReAct paradigm. **Bottom:** The two-dimensional evaluation protocol, requiring adherence to both commonsense rules and hard constraints to achieve a **Success**.

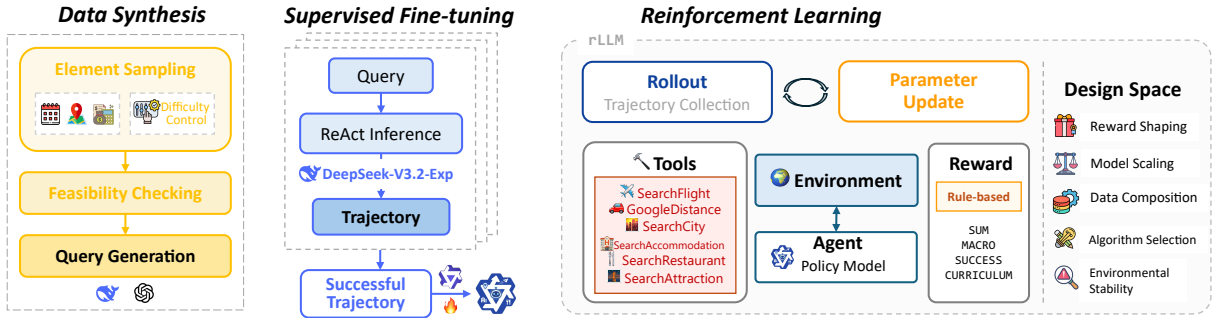


Figure 3: **Overview of the STAR pipeline and the experimental design space.** The pipeline consists of three stages: (1) **Data Synthesis** for generating feasible, difficulty-controlled queries, (2) **SFT** using high-quality trajectories from a teacher model, and (3) **RL** to optimize long-horizon planning via environmental feedback. The right panel outlines the design space explored in our subsequent experiments.

facilitate multi-turn agentic workflows. Given a natural language query q specifying the travel intent and constraints, the agent engages in iterative cycles of reasoning, acting, and observing. At each time step t , the LLM generates a reasoning trace τ_t conditioned on the context, emits a parsable tool action a_t , and receives an observation o_t from the testbed. The process terminates when the agent produces a final natural language itinerary a_T , yielding a complete trajectory defined as:

$$\mathcal{H}_T = (q, \tau_1, a_1, o_1, \dots, \tau_T, a_T).$$

Evaluation Protocol: Given the unstructured nature of the final natural language plan a_T , we employ a dedicated formatting model to parse the output into a structured JSON itinerary prior to automated evaluation, as detailed in Appendix B.4. We evaluate performance along two dimensions, with specific rules outlined in Table 8: **Commonsense** (denoted as cs, e.g., logical consistency, absence of hallucinations) and **Hard Constraints** (denoted as hard, e.g., adherence to budget and dietary restrictions). For each dimension $k \in \{\text{cs}, \text{hard}\}$, we

compute a **micro score** $s_k^{\text{micro}} \in [0, 1]$, representing the ratio of satisfied checks, and a binary **macro score** $s_k^{\text{macro}} = \mathbb{I}(s_k^{\text{micro}} = 1)$, indicating full compliance. A trajectory is deemed **Success** if and only if all constraints are met as follows:

$$s^{\text{success}} = \mathbb{I}(s_{\text{cs}}^{\text{macro}} = 1 \wedge s_{\text{hard}}^{\text{macro}} = 1).$$

3 STAR Pipeline

We instantiate the **STAR** pipeline, a unified post-training framework for controlled study of long-horizon agents on TravelPlanner. As illustrated in Figure 3, the pipeline comprises three sequential stages: data synthesis to construct queries with controllable difficulty, SFT to obtain task-aware initial policies, and RL to further strengthen long-horizon planning behaviors.

Data Synthesis: Addressing the scarcity of training data, we develop a synthesis procedure to generate additional TravelPlanner-style queries. We first sample atomic travel elements (e.g., origin, destination, and dates) and validate their feasibility within the sandbox to ensure the existence of ground-truth solutions. Using these validated constraints

and dynamically estimated budgets, we employ open-sourced models (DeepSeek-AI, 2025; OpenAI, 2025a) to generate natural language queries via back-translation. To obtain queries with controllable difficulty, we follow the original TravelPlanner design, categorizing them into specific difficulty levels (i.e., easy, medium, and hard) by varying the number and types of constraints. Detailed definitions and concrete examples of difficulty levels are provided in **Appendix Table 7**.

SFT: To mitigate the cold-start issue and equip the policy with basic task understanding, we apply SFT prior to RL. We follow a rejection-sampling style procedure: first selecting a strong teacher model to perform ReAct inference on the synthesized queries, retaining only trajectories that achieve *Success* under the evaluation protocol. The resulting high-quality trajectories serve as gold supervision for SFT, yielding task-specialized initial checkpoints for all model sizes.

RL: The core of our framework is the RL stage, where the agent optimizes long-horizon planning via environmental feedback. We utilize rLLM (Tan et al., 2025), a popular framework for post-training language agents. Aligned with the evaluation protocol, we implement a spectrum of reward signals ranging from dense to sparse:

- **SUM:** A dense reward aggregating all sub-metrics, defined as $r^{\text{sum}} = s_{\text{cs}}^{\text{micro}} + s_{\text{cs}}^{\text{macro}} + s_{\text{hard}}^{\text{micro}} + s_{\text{hard}}^{\text{macro}} + s^{\text{success}}$.
- **MACRO:** A semi-sparse reward focusing on macro-level constraint satisfaction, defined as $r^{\text{macro}} = s_{\text{cs}}^{\text{macro}} + s_{\text{hard}}^{\text{macro}} + s^{\text{success}}$.
- **SUCCESS:** A purely sparse binary reward, defined as $r^{\text{success}} = s^{\text{success}}$.
- **CURRICULUM:** Following Zhu et al. (2025), we implement a staged curriculum where the reward function transitions from r^{sum} to r^{success} during training to guide exploration.

We employ GRPO (Shao et al., 2024) as the primary optimization algorithm. For a query q , we sample a group of G trajectories $\{\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(G)}\}$ from the old policy $\pi_{\theta_{\text{old}}}$. The objective maximizes the surrogate advantage as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[\frac{1}{\sum_{i=1}^G |\mathcal{H}^{(i)}|} \sum_{i=1}^G \sum_{j=1}^{|\mathcal{H}^{(i)}|} \min \left(\rho_j^{(i)} \hat{A}^{(i)}, \text{clip}(\rho_j^{(i)}, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) \hat{A}^{(i)} \right) \right],$$

where $\rho_j^{(i)}$ is the importance sampling ratio, ϵ_{low} and ϵ_{high} denote the asymmetric clipping bounds, and $\hat{A}^{(i)}$ is the advantage computed by normalizing rewards within the sampled group. Finally, to systematically explore the RL design space, we extend rLLM into a modular setup that flexibly varies data, rewards, algorithms, and environmental dynamics, facilitating subsequent empirical study.

4 Experiments

4.1 Setup

Pipeline Instantiation: We instantiate the three-stage STAR pipeline with strict quality controls to ensure a rigorous testbed.

- **Data Synthesis:** We synthesize over 10K queries with a balanced difficulty ratio using strong open-weight models, including GPT-OSS-120B (OpenAI, 2025a) and DeepSeek-V3.2-Exp (DeepSeek-AI, 2025). To verify data reliability, we evaluate 200 sampled synthetic queries and confirm that their success rate closely aligns with that of the official TravelPlanner validation set.
- **SFT:** We prompt DeepSeek-V3.2-Exp-Thinking on 5K synthetic queries to perform ReAct inference. Filtering strictly for task *Success* and format adherence yields **1,198 high-quality trajectories** that average 10.3K tokens and 9.2 tool calls, as detailed in Appendix Table 9. We use these to fine-tune the **Qwen2.5-Instruct** series (Qwen et al., 2025) as our SFT base. We intentionally restrict the scale of this SFT phase to establish protocol adherence without inducing policy collapse, thereby preserving exploration space for the subsequent RL stage.
- **RL:** We employ GRPO with practical modifications following Yu et al. (2025a) to stabilize training: (1) KL-Free & Clip-high: We remove the KL penalty and increase the clipping bound ϵ_{high} to encourage broader exploration. (2) Strict protocol enforcement: Trajectories with format errors, e.g., missing `<think> </think>` tags, receive a reward of 0. (3) Overlength handling: To prevent instability, overlength rollouts are excluded from loss computation but retained for advantage normalization to maintain statistical robustness, following Zhao et al. (2026).

Table 1: **In-domain performance (%) across reward designs on the TravelPlanner test set.** Best results for each model scale are in **bold**. Smaller models benefit most from the progressive CURRICULUM reward, while the stronger 7B model favors the dense SUM reward.

Scale	Method	Commonsense		Hard Constraint		Success
		Micro	Macro	Micro	Macro	
1.5B	Base	30.1	0.0	0.0	0.0	0.0
	SFT	65.9	15.4	17.2	12.1	6.9
	SUM	95.1	71.6	51.4	33.4	33.1
	MACRO	93.4	68.4	47.9	31.6	30.1
	SUCCESS	93.9	68.2	51.0	34.7	33.8
	CURRICULUM	93.9	70.2	51.0	35.4	34.9
3B	Base	46.5	0.0	0.0	0.0	0.0
	SFT	70.4	24.6	28.6	20.0	12.2
	SUM	97.6	82.5	64.8	47.6	47.0
	MACRO	95.1	79.6	68.8	52.3	48.2
	SUCCESS	90.6	76.1	62.6	47.7	46.6
	CURRICULUM	95.3	83.0	67.6	51.0	49.9
7B	Base	55.8	0.1	0.7	0.7	0.1
	SFT	77.1	33.1	40.6	31.3	19.7
	SUM	96.9	87.4	78.7	66.5	62.8
	MACRO	97.0	89.8	77.3	60.0	58.9
	SUCCESS	91.4	78.3	73.4	64.0	60.9
	CURRICULUM	96.1	85.1	76.6	60.4	57.0

Default Configurations: Unless otherwise specified, our default RL training uses **1K synthetic queries**, ensuring no overlap with the SFT data, with a 4:3:3 (easy:medium:hard) difficulty ratio. Models are trained for 5 epochs with a group size $G = 8$. The maximum context length is set to 30K tokens during training and extended to 32K for inference. Model selection relies on the TravelPlanner validation set. We conduct controlled experiments by strictly varying one factor at a time while keeping others fixed.

Evaluation: We evaluate in-domain performance on the 1,000-instance TravelPlanner test set and OOD generalization in two settings. First, we report results on 7 knowledge-intensive QA benchmarks using a local Wikipedia search engine following Jin et al. (2025b). Second, we evaluate our 7B checkpoints on BrowseComp-Plus (Chen et al., 2025), a deep-research-style information-seeking benchmark. We report answer accuracy, evidence recall, and average search actions, and include Search-R1-32B (Jin et al., 2025b) under the same retriever setup. Further details are provided in Appendix C.

4.2 Reward Shaping

Motivation: A critical open question in RL for long-horizon agents is how the density of reward signals impacts reasoning capabilities. To answer this question, we evaluate a spectrum of reward

Table 2: **OOD performance on BrowseComp-Plus.** Best results among our 7B variants are in **bold**.

Method	Accuracy (%)	Evidence Recall (%)	Avg. Search
Base-7B	3.5	7.4	1.1
SFT-7B	7.4	14.0	13.4
SUM-7B	6.0	12.6	16.3
MACRO-7B	11.2	15.3	9.0
SUCCESS-7B	9.9	14.2	11.9
CURRICULUM-7B	8.4	13.6	14.8
Search-R1-32B	10.4	10.2	1.7

designs ranging from dense SUM and semi-sparse MACRO, to purely sparse SUCCESS. Furthermore, we evaluate a CURRICULUM reward (Zhu et al., 2025) that progressively transitions from dense to sparse. This acts as a staged intervention based on human priors, providing fine-grained guidance during the early training phases. To strictly isolate the effect of reward shaping, all RL configurations share identical training data, base models, and hyperparameters, as detailed in Appendix D.1.

Table 1 presents the in-domain performance on TravelPlanner. We compare our RL variants against two baselines: the pre-trained Base models and their SFT counterparts, which serve as the starting checkpoints for RL. For a comprehensive analysis, training dynamics and OOD results on the 7 knowledge-intensive QA benchmarks are provided in Figure 8 and Table 10 in Appendix D.1, respectively. To further examine transfer to a more challenging information-seeking setting, Table 2 reports the performance of our 7B variants on BrowseComp-Plus. Synthesizing these results yields two key takeaways.

Takeaway 1: Reward design is scale-dependent; sparse-only rewards are suboptimal. In the TravelPlanner domain, smaller models struggle with credit assignment over long horizons and benefit significantly from staged guidance. Consequently, the CURRICULUM reward achieves the highest success rates and accelerates convergence, as shown in Figure 8. Conversely, the stronger 7B model possesses the intrinsic capacity to directly leverage fine-grained feedback from the dense SUM reward, rendering heuristic staged transitions unnecessary and even slightly restrictive. Notably, while the sparse SUCCESS reward is competitive, it never achieves the best performance across any scale, indicating that outcome-only feedback is insufficient for optimizing long-horizon trajectories.

Takeaway 2: Dense task-specific rewards can induce an alignment tax on OOD generalization. While the SUM reward maximizes in-domain

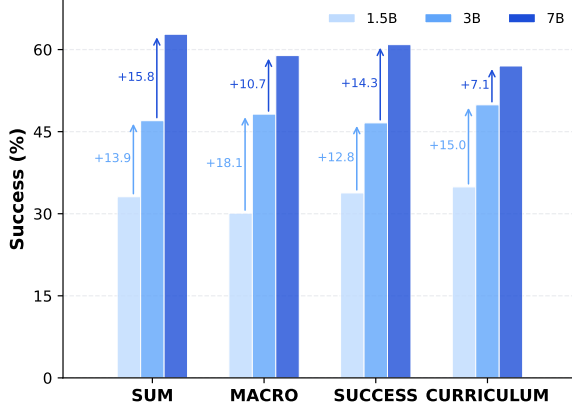


Figure 4: **Performance across model scales on the TravelPlanner test set.** Scaling up model capacity consistently improves success rates across all reward types, though the magnitude of improvement varies.

performance for the 7B model, its average accuracy on the knowledge-intensive QA benchmarks drops from 41.9% at the SFT checkpoint to 36.7%, as shown in Table 10. A similar pattern emerges on BrowseComp-Plus in Table 2: MACRO achieves the highest answer accuracy and evidence recall among our 7B variants, while requiring fewer search actions than the other RL variants. It also exceeds Search-R1-32B in both accuracy and evidence recall under the same retriever setup. These results suggest that optimizing overly dense, task-specific feedback can sacrifice transferable information-seeking behavior, whereas the semi-sparse MACRO reward achieves a better balance between in-domain optimization and OOD transfer.

4.3 Model Scaling

Motivation: Beyond reward design, a fundamental question is whether scaling model capacity inherently resolves the reasoning bottlenecks in long-horizon RL. To investigate this, we compare the 1.5B, 3B, and 7B models under fixed reward configurations. This allows us to evaluate if larger underlying architectures are better equipped to handle the complexities of multi-turn tool use and planning. Figure 4 illustrates the in-domain success rates on the TravelPlanner test set across different model scales. Synthesizing these results, along with the training dynamics shown in Figure 9 in Appendix D.2, reveals a clear scaling behavior.

Takeaway 3: Scaling model size consistently improves long-horizon performance, but the magnitude of gains varies across reward designs. As shown in Figure 4, transitioning from the 1.5B

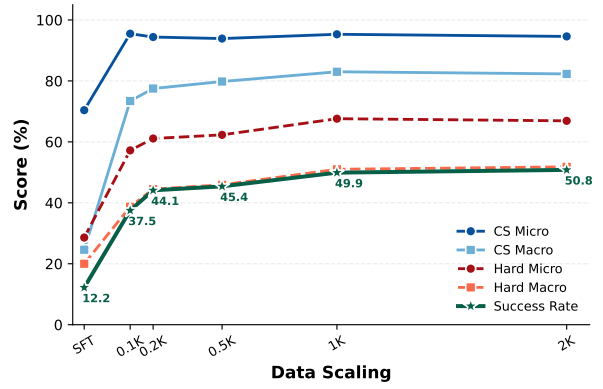


Figure 5: **Performance across different training data quantities on the TravelPlanner test set.** A sweet spot emerges at 1K prompts, balancing in-domain success with strong OOD generalization, which is detailed in Table 11 in Appendix D.3.

to the 7B architecture yields substantial improvements in success rates across all reward signals. For instance, under the dense SUM reward, the success rate nearly doubles from 33.1% at 1.5B to 62.8% at 7B. This upward trend is further corroborated by the training dynamics in Figure 9, which demonstrate that larger models not only converge faster but also reach significantly higher performance asymptotes. While scaling is universally beneficial, we observe that the specific rate of improvement is reward-dependent, e.g., moving from 3B to 7B yields a 15.8% absolute gain under SUM, vs. only 7.1% under CURRICULUM. Ultimately, these findings indicate that base model capacity remains a primary bottleneck for complex agentic tasks, and that RL effectively unlocks these inherent reasoning capabilities, particularly when guided by suitable reward designs.

4.4 Data Composition

Motivation: While SFT typically benefits from massive data volumes, the optimal data strategy for RL in complex agentic tasks remains under-explored. We investigate RL data composition across two orthogonal dimensions: *quantity* and *difficulty*. For quantity, we ask whether RL exhibits a continuous scaling law or a saturation point where over-optimization degrades generalization. For difficulty, we examine how the mixture of task complexity influences the resulting planning capabilities. Building upon our findings in previous sections, we fix the base model at 3B and utilize the CURRICULUM reward, the optimal configuration for models of this scale, to strictly isolate these

Table 3: **In-domain performance (%) across data difficulty compositions on the TravelPlanner test set.** Models are trained on 1K prompts from the same SFT-3B checkpoint using the CURRICULUM reward. The notation XX-1K denotes the data composition, e.g., Easy-1K contains exclusively easy queries. **Mixed-1K** blends easy:medium:hard at a 4:3:3 ratio.

Data	Commonsense		Hard Constraint		Success
	Micro	Macro	Micro	Macro	
Easy-1K	92.8	79.7	62.8	46.9	45.3
Medium-1K	94.9	69.2	66.2	49.0	41.1
Hard-1K	90.0	48.4	65.5	47.2	25.9
Mixed-1K	95.3	83.0	67.6	51.0	49.9

data variables. Detailed experimental setups, training dynamics analysis, and full OOD results are deferred to Appendix D.3.

Takeaway 4: RL data scaling exhibits a sweet spot. As illustrated in Figure 5, increasing the training data from 100 to 1K prompts yields a rapid improvement in the in-domain success rate, rising from 37.5% to 49.9%. Concurrently, the average OOD score, shown in Table 11, reaches its peak at 35.0%. Scaling further to 2K prompts provides only a marginal in-domain gain to 50.8%, while the OOD score drops to 32.2%. These results suggest that, in our setting, 1K prompts provide a strong balance between in-domain optimization and OOD generalization.

Takeaway 5: Balanced data difficulty prevents reward sparsity and enables complex constraint mastery. Table 3 compares models trained on varying difficulty levels, discriminated by the number of constraints. For example, **easy** samples typically contain only a single budget limit, whereas **hard** samples introduce compounding requirements across transportation, meals, and accommodation. Training exclusively on easy data allows the model to grasp basic planning, achieving a high Commonsense Macro score of 79.7%, but fails to teach complex constraint satisfaction. Conversely, training solely on hard data leads to substantial performance degradation. The multifaceted constraints make successful trajectories exceedingly rare, exacerbating reward sparsity and destabilizing policy learning. The mixed configuration effectively resolves this dilemma. By blending difficulty levels, it provides enough simple tasks to maintain dense reward signals for commonsense learning, while incorporating sufficient complex tasks to teach advanced constraint satisfaction, ultimately achieving the highest overall success rate of 49.9%.

4.5 Algorithm Selection

Motivation: Recent advancements in agentic RL often introduce sophisticated sampling mechanisms to encourage exploration. To determine whether training long-horizon agents requires such algorithmic designs, we benchmark the standard GRPO against two representative variants: DAPO (Yu et al., 2025a) and ARPO (Dong et al., 2026). DAPO represents reward-guided trajectory filtering, e.g., discarding batches with zero variance in rewards, while ARPO represents adaptive rollout mechanisms that utilize entropy to dynamically branch trajectories. To provide appropriate reward signals across scales, we apply the MACRO reward for the 1.5B and 3B models, and the SUM reward for the 7B model. To ensure a fair comparison, all algorithms share identical technical enhancements, with the key distinction lying solely in their exploration strategies. More implementation details and training dynamics are deferred to Appendix D.4.

Table 4 presents a comprehensive comparison of both effectiveness and training efficiency across model scales, leading to the following insight.

Takeaway 6: The necessity for sophisticated exploration is inversely correlated with model scale. As shown in Table 4, smaller models inherently struggle with sparse reward signals. At the 1.5B scale, algorithmic interventions like ARPO and DAPO significantly outperform GRPO, achieving success rates of 37.5% and 36.9%, respectively, compared to 30.1%. This advantage is also visually evident in their training dynamics in Appendix Figure 11, where exploration-heavy algorithms demonstrate faster convergence and higher peak performance early in training. However, as model capacity increases to 3B, this performance gap closes. Strikingly, at the 7B scale, GRPO achieves the highest success rate of 62.8%, outperforming both DAPO and ARPO. Furthermore, sophisticated exploration introduces extra computational overhead. For instance, ARPO’s adaptive rollout requires calculating the entropy difference at each generation step to determine whether to branch new trajectories. While intuitive, this mechanism increases the time cost per step. Therefore, these trends highlight a practical paradigm shift for scaling agentic RL: rather than expending computational resources and engineering effort on complex heuristic algorithms, we can leverage the inherent exploration capabilities of stronger base models, relying on the raw efficiency of GRPO algorithm.

Table 4: **Performance (%) and training efficiency across different RL algorithms on the TravelPlanner test set.** Best results for each metric within the same model scale are in **bold**. GPU Hours are calculated based on $8 \times A100-80G$ for 1.5B and 3B models and $16 \times A100$ for the 7B model.

Scale	Algorithm	Commonsense		Hard Constraint		Success $\uparrow$	Time / Step (min) $\downarrow$	GPU Hours $\downarrow$
		Micro	Macro	Micro	Macro			
1.5B	GRPO	93.4	68.4	47.9	31.6	30.1	8.0	164
	DAPO	94.4	77.7	55.6	38.2	36.9	8.2	183
	ARPO	93.1	76.4	58.1	39.4	37.5	9.5	195
3B	GRPO	95.1	79.6	68.8	52.3	48.2	8.6	176
	DAPO	93.6	78.6	66.1	48.7	45.6	8.9	184
	ARPO	95.4	80.3	67.9	51.5	47.5	13.3	273
7B	GRPO	96.9	87.4	78.7	66.5	62.8	9.0	368
	DAPO	95.6	87.9	75.4	60.6	58.4	9.5	390
	ARPO	96.7	86.8	76.6	61.1	58.3	13.3	547

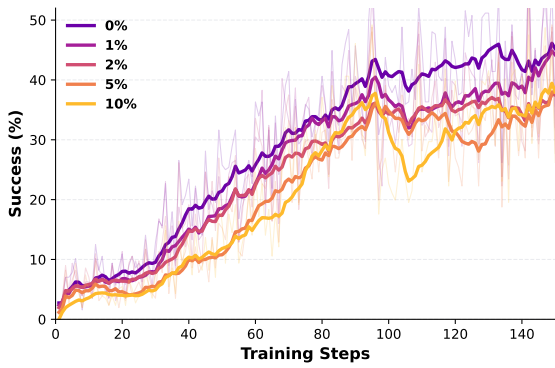


Figure 6: **Training dynamics under randomly injected API dropout.** During RL training, each tool execution independently returns a tool unavailable error with the probability shown in the legend. The curves report success on the training rollouts at each step. High dropout rates slow convergence and increase training variance.

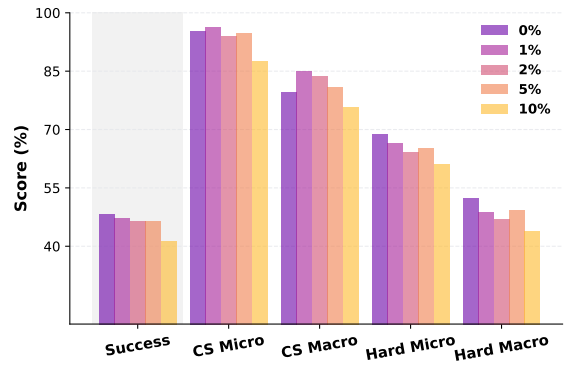


Figure 7: **Clean-test performance after training with randomly injected API dropout.** The percentages denote the API dropout probabilities applied *only during RL training*; all checkpoints are evaluated on the same clean, failure-free TravelPlanner test set. Performance remains relatively stable under mild dropout but degrades noticeably at high training-time instability.

4.6 Environmental Stability

Motivation: Agentic RL relies on repeated interactions with the environment during rollout, yet these interactions may be unreliable in practice. We therefore ask whether *training-time* environmental instability can impair policy learning even when the resulting agent is evaluated in a clean environment. We first study **API Dropout**, where each tool execution is independently replaced with a tool unavailable response with a predefined probability. We vary this dropout probability from 0% to 10% during RL training, while keeping the reward design (MACRO) and all other hyperparameters fixed. Importantly, all resulting checkpoints are evaluated on the same clean, failure-free TravelPlanner test environment, isolating the effect of environmental instability on policy optimization rather than test-time execution.

Figures 6 and 7 show the resulting training dynam-

ics and clean-test performance, respectively. To further examine whether the *type* of environmental noise matters, we additionally introduce two forms of randomly injected observation corruption during training: **Missing-field**, where information in tool-returned records is randomly masked, and **Missing + Distractor**, which further injects irrelevant records into the observations. Table 5 compares these perturbations with API Dropout at $p = 0.05$ and $p = 0.10$, again evaluating all trained policies in the clean environment.

Takeaway 7: Training-time environmental instability can degrade the learned policy, and observation corruption is more harmful than explicit API failures. Figures 6 and 7 show that the agent tolerates mild API dropout during training, whereas higher dropout rates slow convergence and reduce subsequent performance even when evaluation is fully clean. The type of instability also matters.

Table 5: **Impact of randomly injected environmental perturbations during RL training.** All 3B models use the MACRO reward. Perturbations are applied only during training, and all resulting checkpoints are evaluated on the same clean TravelPlanner test environment.

Perturbation	p	Commonsense		Hard Constraint		Success
		Micro	Macro	Micro	Macro	
No Errors	–	95.1	79.6	68.8	52.3	48.2
API Dropout	0.05	94.8	80.8	65.2	49.1	46.4
Missing-field	0.05	93.2	75.2	64.1	47.5	44.4
Missing + Distractor	0.05	91.3	71.8	58.9	44.6	39.5
API Dropout	0.10	87.4	75.7	61.1	43.7	41.3
Missing-field	0.10	90.5	71.4	57.5	40.3	36.8
Missing + Distractor	0.10	93.6	67.8	53.1	34.9	33.2

As shown in Table 5, at $p = 0.05$, API Dropout reduces success to 46.4%, whereas Missing-field and Missing + Distractor reduce it further to 44.4% and 39.5%, respectively. The same ordering persists at $p = 0.10$. We hypothesize that explicit API failures are easier for the agent to recognize, while partially missing or misleading observations allow it to continue planning from unreliable evidence. These results highlight the importance of reliable rollout environments for stable long-horizon RL training.

5 Conclusion

In this paper, we present a systematic empirical study to demystify the RL design space for long-horizon, tool-using agents. Using TravelPlanner as a testbed, we implement the flexible STAR pipeline to decompose the post-training process along 5 axes: reward shaping, model scaling, data composition, algorithm selection, and environmental stability. Our extensive empirical study yields a practical, scale-aware recipe: smaller models require staged curriculum rewards and sophisticated exploration, whereas larger models learn efficiently with simpler dense rewards and standard GRPO. Furthermore, we identified a data sweet spot in quantity and difficulty that balances in-domain mastery and OOD generalization. Applying this recipe, STAR achieves a new SOTA on TravelPlanner with much smaller models, significantly outperforming leading proprietary LLMs.

Our findings provide actionable guidelines for training long-horizon agents and highlight open challenges, e.g., how multiple factors simultaneously guide policy learning. We hope STAR serves as a valuable toolkit to inspire future research.

Acknowledgements

This research is supported by grants from the Research Grants Council of the Hong Kong SAR, China (No. CUHK 14206625).

Limitations

Our study is intentionally designed as a controlled empirical analysis of long-horizon agentic RL. Several questions nevertheless remain open.

- (1) **From controlled simulation to live environments:** We deliberately use TravelPlanner as a controlled testbed: its local sandbox provides deterministic, low-cost, and high-throughput interactions, making large-scale RL ablations reproducible while retaining multi-tool orchestration and constraint-satisfaction complexity grounded in real-world records. We further stress-test this setting with explicit tool failures and corrupted observations. However, live APIs introduce additional dynamics, including latency, rate limits, schema drift, and non-deterministic responses across runs. Whether the empirical recipes identified here transfer unchanged to such production environments remains an open question.
- (2) **Breadth of OOD evaluation:** Our transfer evaluation covers knowledge-intensive QA and deep-research-style information seeking, providing evidence beyond the TravelPlanner domain. However, these benchmarks remain primarily information-seeking tasks. Domains with fundamentally different action spaces and interaction structures, such as software engineering agents, GUI control, and richer multi-tool environments, remain untested.
- (3) **Scaling to substantially larger models:** Due to computational constraints, our experiments cover models up to 7B parameters. Whether the scale-dependent trends observed in this range continue to hold for substantially larger LLMs remains an open question.
- (4) **Single-factor analysis and task-specific rewards:** To enable controlled comparisons, we primarily vary one design axis at a time, leaving interactions among reward shaping, model scale, data composition, algorithm choice, and environmental stability largely unexplored. In addition, our reward functions are derived from TravelPlanner’s trajectory-level evaluation signals. Studying more task-agnostic reward formulations and finer-grained credit assignment may further clarify how to address reward sparsity in long-horizon settings.

(5) **Global re-planning and backtracking:** An orthogonal challenge, shared by current long-horizon agents more broadly, is global re-planning after an upstream decision becomes infeasible. Our case analysis shows that local adaptation can emerge, while revising earlier planning decisions remains difficult. In our setup, this may stem from the linear ReAct interaction paradigm, which does not explicitly maintain alternative plans or rollback states, together with trajectory-level rewards that provide only indirect supervision for *when* to backtrack. Addressing this capability may require explicit plan-state tracking, search-based mechanisms, or finer-grained credit assignment, and lies outside the RL design axes studied in this work.

References

- Moonshot AI. 2025. [Kimi-researcher: End-to-end rl training for emerging agentic capabilities](https://moonshotai.github.io/Kimi-Researcher/). <https://moonshotai.github.io/Kimi-Researcher/>.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhui Chen, and Jimmy Lin. 2025. [Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent](#). *arXiv preprint arXiv:2508.06600*.
- Xiang Cheng, Yulan Hu, Xiangwen Zhang, Lu Xu, Lide Tan, Zheng Pan, Xin Li, and Yong Liu. 2026. [Beyond itinerary planning-a real-world benchmark for multi-turn and tool-using travel tasks](#). *Preprint*, arXiv:2512.22673.
- DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.
- Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025. [Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning](#). *Preprint*, arXiv:2505.16410.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazheng Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2026. [Agentic reinforced policy optimization](#). In *The 14th International Conference on Learning Representations*.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. [Group-in-group policy optimization for LLM agent training](#). In *The 39th Annual Conference on Neural Information Processing Systems*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. 2026. [Tree search for LLM agent reinforcement learning](#). In *The 14th International Conference on Learning Representations*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. [SWE-bench: Can language models resolve real-world github issues?](#) In *The 12th International Conference on Learning Representations (ICLR)*.
- Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O Arik, and Jiawei Han. 2025a. [An empirical study on reinforcement learning for reasoning-search interleaved llm agents](#). *arXiv preprint arXiv:2505.15117*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025b. [Search-rl: Training LLMs to reason and leverage search engines with reinforcement learning](#). In *Second Conference on Language Modeling*.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). *Preprint*, arXiv:1705.03551.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Ding-Chu Zhang, Xixi Wu, Xinmiao Yu, Jialong Wu, Xinyu Wang, Zile Qiao, Zhen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Zhi-Qin John Xu, Shuai Wang, and 2 others. 2026a. [Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning](#). In *The 14th International Conference on Learning Representations*.
- Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. 2025. [Websailor: Navigating super-human reasoning for web agent](#). *arXiv preprint arXiv:2507.02592*.
- Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, Jun Zhang, and Jingren Zhou. 2026b.

- Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. In *The 14th International Conference on Learning Representations*.
- Zhihan Liu, Lin Guan, Yixin Nie, Kai Zhang, Zhuoqun Hao, Lin Chen, Asli Celikyilmaz, Zhaoran Wang, and Na Zhang. 2026. [Paying less generalization tax: A cross-domain generalization study of rl training for llm agents](#). *Preprint*, arXiv:2601.18217.
- Siyuan Lu, Zechuan Wang, Hongxuan Zhang, Qintong Wu, Leilei Gan, Chenyi Zhuang, Jinjie Gu, and Tao Lin. 2026. [Don't just fine-tune the agent, tune the environment](#). In *The 14th International Conference on Learning Representations*.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). *Preprint*, arXiv:2212.10511.
- Yansong Ning, Rui Liu, Jun Wang, Kai Chen, Wei Li, Jun Fang, Kan Zheng, Naiqiang Tan, and Hao Liu. 2025. [Deepravel: An end-to-end agentic reinforcement learning framework for autonomous travel planning agents](#). *Preprint*, arXiv:2509.21842.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025a. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- OpenAI. 2025b. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. [Training language models to follow instructions with human feedback](#). *ArXiv*, abs/2203.02155.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Bytedance Seed. 2026. [Seed1.8 model card: Towards generalized real-world agency](#). *Preprint*, arXiv:2603.20633.
- Jie-Jing Shao, Bo-Wen Zhang, Xiao-Wen Yang, Baizhi Chen, Siyu Han, Pang Jinghao, Wen-Da Wei, Guohao Cai, Zhenhua Dong, Lan-Zhe Guo, and Yu-Feng Li. 2026. [Chinatravel: An open-ended travel planning benchmark with compositional constraint validation for language agents](#). In *The 14th International Conference on Learning Representations*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Yuanzhe Shen, Zisu Huang, Zhengyuan Wang, Muzhao Tian, Zhengkang Guo, Chenyang Zhang, Shuaiyu Zhou, Zengjie Hu, Dailin Li, Jingwen Xu, Kaimin Wang, Wenhao Liu, Tianlong Li, Fengpeng Yue, Feng Hong, Cao Liu, and Ke Zeng. 2026. [Trip-bench: A benchmark for long-horizon interactive agents in real-world scenarios](#). *Preprint*, arXiv:2602.01675.
- Sijun Tan, Michael Luo, Colin Cai, Tarun Venkat, Kyle Montgomery, Aaron Hao, Tianhao Wu, Arnav Balyan, Manan Roongta, Chenguang Wang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. [rllm: A framework for post-training language agents](#).
- Zhengwei Tao, Jialong Wu, Wenbiao Yin, Pu Wu, Junkai Zhang, Baixuan Li, Haiyang SHEN, Kuan Li, Liwen Zhang, Xinyu Wang, Wentao Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2026. [Webshaper: Agentic data synthesizing via information-seeking formalization](#). In *The 14th International Conference on Learning Representations*.
- Gemini Team. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Kimi Team. 2026a. [Kimi k2.5: Visual agentic intelligence](#). *Preprint*, arXiv:2602.02276.
- Qwen Team. 2026b. [Qwen3.5: Towards native multimodal agents](#).
- Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. 2025. [Tongyi deepresearch technical report](#). *arXiv preprint arXiv:2510.24701*.
- Harsh Trivedi, Niranjana Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [MuSiQue: Multi-hop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.

- Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, Wanjun Zhong, Yining Ye, Yujia Qin, Yuwen Xiong, Yuxin Song, Zhiyong Wu, Aoyan Li, Bo Li, Chen Dun, and 93 others. 2025. [Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning](#). *Preprint*, arXiv:2509.02544.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxiang Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. [Text embeddings by weakly-supervised contrastive pre-training](#). *Preprint*, arXiv:2212.03533.
- Xiaoxuan Wang, Han Zhang, Haixin Wang, Yidan Shi, Ruoyan Li, Kaiqiao Han, Chenyi Tong, Haoran Deng, Alexander K Taylor, Renliang Sun, Yanqiao Zhu, Jason Cong, Yizhou Sun, and Wei Wang. 2026a. [AR-LArena: A unified framework for stable agentic reinforcement learning](#). In *Forty-third International Conference on Machine Learning*.
- Yinjie Wang, Tianbao Xie, Ke Shen, Mengdi Wang, and Ling Yang. 2026b. [RLAnything: Forge environment, policy, and reward model in completely dynamic RL system](#). In *Forty-third International Conference on Machine Learning*.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Xiangru Tang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. [Webdancer: Towards autonomous information seeking agency](#). In *The 39th Annual Conference on Neural Information Processing Systems*.
- Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Minhao Cheng, Shuai Wang, Hong Cheng, and Jingren Zhou. 2025b. [Resum: Unlocking long-horizon search intelligence via context summarization](#). *arXiv preprint arXiv:2509.13313*.
- Xixi Wu, Yifei Shen, Caihua Shan, Kaitao Song, Siwei Wang, Bohang Zhang, Jiarui Feng, Hong Cheng, Wei Chen, Yun Xiong, and Dongsheng Li. 2024. [Can graph learning improve planning in llm-based agents?](#) In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*.
- Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, yajie yang, Rui Zheng, Junjie Ye, Jiazhen Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, and 5 others. 2026. [Agentgym-RL: An open-source framework to train LLM agents for long-horizon decision making via multi-turn RL](#). In *The 14th International Conference on Learning Representations*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024a. [Travelplanner: A benchmark for real-world planning with language agents](#). In *Forty-first International Conference on Machine Learning*.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024b. [Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments](#). In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*.
- John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. 2024. [SWE-agent: Agent-computer interfaces enable automated software engineering](#). In *The 38th Conference on Neural Information Processing Systems (NeurIPS)*.
- John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. 2025a. [SWE-bench multimodal: Do ai systems generalize to visual software domains?](#) In *The 13th International Conference on Learning Representations (ICLR)*.
- Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. 2025b. [Demystifying long chain-of-thought reasoning](#). In *Forty-second International Conference on Machine Learning*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The 11th International Conference on Learning Representations (ICLR)*.
- Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiyong Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2026. [Memagent: Reshaping long-context LLM with multi-conv RL-based memory agent](#). In *The 14th International Conference on Learning Representations*.
- Qiyong Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025a. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *The 39th Annual Conference on Neural Information Processing Systems*.

- Zhaochen Yu, Ling Yang, Jiaru Zou, Shuicheng Yan, and Mengdi Wang. 2025b. [Demystifying reinforcement learning in agentic reasoning](#). *arXiv preprint arXiv:2510.11701*.
- Ruiyang Zhang, Qianguo Sun, Chao Song, Yiyang Qi, and Zhedong Zheng. 2026a. [Vsearcher: Long-horizon multimodal search agent via reinforcement learning](#). *Preprint*, arXiv:2603.02795.
- Siwei Zhang, Yun Xiong, Xi Chen, Zi'an Jia, Renhong Huang, Jiarong Xu, and Jiawei Zhang. 2026b. [Rapo: Expanding exploration for llm agents via retrieval-augmented policy optimization](#). In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Yinger Zhang, Shutong Jiang, Renhao Li, Jianhong Tu, Yang Su, Lianghao Deng, Xudong Guo, Chenxu Lv, and Junyang Lin. 2026c. [Deepplanning: Benchmarking long-horizon agentic planning with verifiable constraints](#). *Preprint*, arXiv:2601.18137.
- Yida Zhao, Kuan Li, Xixi Wu, Liwen Zhang, Ding-Chu Zhang, Baixuan Li, Maojia Song, Zhuo Chen, Chenxi Wang, Xinyu Wang, Yong Jiang, Kewei Tu, Pengjun Xie, Fei Huang, and Jingren Zhou. 2026. [Repurposing synthetic data for fine-grained search agent supervision](#). In *The 14th International Conference on Learning Representations*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. [Group sequence policy optimization](#). *arXiv preprint arXiv:2507.18071*.
- Siyu Zhu, Yanbin Jiang, Hejian Sang, Shao Tang, Qingquan Song, Biao He, Rohit Jain, Zhipeng Wang, and Alborz Geramifard. 2025. [Planner-r1: Reward shaping enables efficient agentic rl with smaller llms](#). *Preprint*, arXiv:2509.25779.

A Related Works

A.1 LLM-based Agents

LLMs have evolved from static text generators into autonomous agents capable of reasoning, acting, and interacting with dynamic environments (Yao et al., 2023; Schick et al., 2023). This evolution has catalyzed applications across diverse domains, spanning deep research agents navigating the open web for information (Li et al., 2025, 2026b; Wu et al., 2025b; Team et al., 2025; Zhang et al., 2026a), software engineering agents debugging real-world codebases (Jimenez et al., 2024; Yang et al., 2025a, 2024), and GUI agents serving as personal assistants (Xie et al., 2024b; Wang et al., 2025).

Among existing agentic environments, we select TravelPlanner (Xie et al., 2024a) as our primary testbed. As a pioneering planning benchmark, alongside recent benchmarks exploring broader planning settings (Zhang et al., 2026c; Shen et al., 2026; Shao et al., 2026; Cheng et al., 2026), it offers critical advantages for our empirical study. First, it features an abundant **1K-instance test set** with **real-world complexity**, demanding multi-tool orchestration and constraint satisfaction. Its inherent **difficulty** is evident, as even advanced models like Kimi-K2.5 (Team, 2026a) achieve a mere 11.8% success rate. Crucially for RL scaling, TravelPlanner provides a local sandbox that enables **zero-cost, high-concurrency** execution. In contrast, the computational overhead of GUI and code agents, coupled with the prohibitive API costs of open-web tasks, creates severe bottlenecks for extensive environmental exploration. These properties make TravelPlanner a scalable and reproducible testbed for studying agentic RL.

A.2 Agentic Reinforcement Learning

While RL from human or AI feedback (Ouyang et al., 2022) has become the standard practice for LLM alignment, applying RL to autonomous agents introduces unique challenges, including complex environmental dynamics, long-horizon requirements, and sparse, delayed reward signals (Xi et al., 2026). To tackle these issues, GRPO (Shao et al., 2024) and its variants (Feng et al., 2025; Zheng et al., 2025; Yu et al., 2025a; Zhang et al., 2026b) have emerged as common algorithms for agentic RL. Building upon this foundation, recent advancements can be categorized into three primary directions: (1) **Paradigm Adaptation**: To han-

dle dynamic environments and long-horizon tasks, researchers are moving beyond the ReAct framework (Yao et al., 2023) to propose novel agentic workflows (Wu et al., 2025b; Yu et al., 2026). This shift has driven the development of specialized RL algorithms tailored to adapt models to these new paradigms. (2) **Exploration-driven Sampling**: To navigate vast action spaces and encourage exploration in highly uncertain states, recent methods modify the sampling process. A representative approach is ARPO (Dong et al., 2026), which adaptively forks new paths upon detecting high entropy during generation. (3) **Reward Shaping for Dense Supervision**: To mitigate the sparsity of outcome-based rewards, intermediate signals are increasingly introduced. For example, E-GRPO (Zhao et al., 2026) allocates partial bonuses for retrieving key entities within the agent’s reasoning traces during deep research tasks. DeepTravel (Ning et al., 2025) employs an LLM-as-a-Judge to provide step-wise, action-level rewards, while Planner-R1 (Zhu et al., 2025) introduces staged reward shaping for travel planning.

Recent agent-training systems such as WebDancer (Wu et al., 2025a), WebSailor (Li et al., 2025), WebShaper (Tao et al., 2026), and ToolStar (Dong et al., 2025) similarly leverage synthesized data and post-training to build stronger web or multi-tool agents. Our work is complementary to these systems: rather than introducing a new post-training pipeline skeleton, we use STAR as a unified scaffold to systematically study how reward shaping, model scale, data composition, algorithm selection, and environmental stability affect long-horizon agentic RL.

A.3 Empirical Studies of Agentic RL

Beyond algorithmic innovations, recent empirical studies seek practical recipes for training LLM agents. For instance, Jin et al. (2025a) analyze reward shaping and model selection, but their scope is limited to short-horizon interactions within the QA domain. While Yu et al. (2025b) explore data composition and policy optimizations, their findings remain confined to mathematical reasoning. Other studies investigate complementary aspects of agentic RL, including training stability (Wang et al., 2026a) and cross-domain generalization (Liu et al., 2026). However, systematic empirical understanding of RL design choices in complex, long-horizon tool-use settings remains limited.

In summary, existing work has made substan-

Table 6: Details of implemented tools within the TravelPlanner testbed.

Tool	Arguments	Description	# Data Entries
SearchCity	state	Find cities within a specific state	64
SearchFlight	departure, destination, date	Search for flight information between cities on specific dates	3,827,361
GoogleDistance	departure, destination, mode	Calculate distance, travel time, and cost between two cities	17,603
SearchRestaurant	city	Find restaurant options in a specific city	9,552
SearchAttraction	city	Find tourist attractions in a specific city	5,303
SearchAccommodation	city	Find accommodation options in a specific city	5,064

Table 7: Examples of queries across different difficulty levels on the TravelPlanner validation set. Different types of constraints are highlighted: **budget**, **accommodation**, **transportation**, **routing**, and **duration**.

Difficulty	Example Query	Description
Easy	Please create a travel plan for me where I’ll be departing from Washington and heading to Myrtle Beach for a 3-day trip from March 13th to March 15th, 2022 . Can you help me keep this journey within a budget of \$1,400 ?	Single people & Basic constraint: Involves planning for 1 person. The only explicit constraint is a budget limit.
Medium	We need to create a travel plan for 2 people departing from Lake Charles and visiting 2 cities in Texas . The trip will last for 5 days , starting from March 4th to March 8th, 2022 . The budget is set at \$4,600 . In terms of accommodations, we prefer places where parties are allowed .	Specific constraints: Requires planning for a group of people. Introduces specific accommodation preferences alongside the budget.
Hard	Could you create a 7-day travel itinerary for 2 people, departing from Albuquerque and visiting 3 cities in Texas from March 8th to March 14th, 2022 ? Our budget is set at \$5,000 . We require accommodations that allow smoking and are preferably not shared rooms . We would prefer to avoid any flights for our transportation.	Multi-dimensional constraints: Demands planning a long-horizon trip. Features highly restrictive, multi-dimensional constraints like strict accommodation rules and transportation limitations.

tial progress in both agent-training systems and individual RL design choices, while practical guidance for configuring RL in long-horizon tool-using settings remains fragmented. Using STAR as a unified experimental scaffold, we conduct controlled comparisons across reward shaping, model scaling, data composition, algorithm selection, and environmental stability, and distill actionable findings from this design space.

B Details of TravelPlanner

In this section, we provide supplementary details regarding the TravelPlanner benchmark (Xie et al., 2024a). This includes the tool execution sandbox, query examples across different difficulty levels, concrete evaluation rules, and the formatting process of the final plan.

B.1 Sandbox

Table 6 outlines the tools supported within the TravelPlanner environment, which enable agents to gather information across various aspects of travel. These tools are implemented by querying a local database, ensuring low latency and zero execution cost. To maintain alignment with the abundance of

real-world information, all database entries were originally collected from real-world APIs by the creators of TravelPlanner, maximizing the reliability and authenticity of the sandbox.

B.2 Instances

The TravelPlanner dataset is divided into three splits: a training set of 45 queries, a validation set of 180 queries, and a test set of 1,000 queries. To illustrate the nature of these queries, we present three examples of varying difficulty levels from the validation set in **Table 7**. Difficulty is primarily determined by the number of constraints. **Easy** queries typically target a single traveler with only a total budget limit. **Medium** queries introduce additional constraints focusing on meals, transportation, or accommodation, often accompanied by an increase in the number of travelers, which complicates budget calculations. Finally, **Hard** queries impose even stricter constraints, presenting challenges for long-horizon planning that must satisfy multi-dimensional requirements simultaneously. Notably, both the difficulty levels and the trip durations are well-balanced across all dataset splits.

Table 8: **Official evaluation rules on TravelPlanner.**

<i>Commonsense</i>	
Within Sandbox	All information in the plan must be within the closed sandbox. Otherwise, it will be considered a hallucination.
Complete Information	No key information should be left out of the plan, such as the lack of accommodation during travel.
Within Current City	All scheduled activities for the day must be located within that day’s city(s).
Reasonable City Route	Changes in cities during the trip must be reasonable.
Diverse Restaurants	Restaurant choices should not be repeated throughout the trip.
Diverse Attractions	Attraction choices should not be repeated throughout the trip.
Non-conflict Transportation	Transportation choices within the trip must be reasonable. For example, having both “self-driving” and “flight” would be considered a conflict.
Minimum Nights Stay	The number of consecutive days spent in a specific accommodation during the trip must meet the corresponding required minimum number of nights’ stay.
<i>Hard Constraint</i>	
Budget	The total budget of the trip.
Room Rule	Room rules include “No parties”, “No smoking”, “No children under 10”, “No pets”, and “No visitors”.
Room Type	Room types include “Entire Room”, “Private Room”, “Shared Room”, and “No Shared Room”.
Cuisine	Cuisines include “Chinese”, “American”, “Italian”, “Mexican”, “Indian”, “Mediterranean”, and “French”.
Transportation	Transportation options include “No flight” and “No self-driving”.

B.3 Evaluation Rules

Following the official evaluation protocol of TravelPlanner, we detail the specific rules for each evaluation dimension in Table 8, categorizing them into commonsense and hard constraints. For each dimension, the micro score is computed as the percentage of satisfied rules. If all rules within a dimension are satisfied, the macro score is set to 1. A generated trajectory is deemed a *Success* only if both commonsense and hard constraints are fully satisfied.

These constraints vividly illustrate the inherent complexity of travel planning. For instance, when selecting accommodations, the agent must satisfy not only commonsense rules, e.g., the minimum stay nights required by a specific hotel, but also the user’s explicitly stated preferences, e.g., room type. Collectively, the agent must ensure that the entire itinerary fits within the overall budget, adheres to transportation rules, remains strictly grounded in the provided context to avoid hallucinations, and forms a logically consistent route.

B.4 Planning Formatting

Following the original evaluation protocol of TravelPlanner, the agent generates travel plans in a natural, creative language format, which we believe aligns with real-world user expectations and readability. However, for the purpose of automated

evaluation, these natural language plans must be converted into a structured JSON format that details distinct daily elements, e.g., transportation, accommodation, and attractions. Rather than forcing the planning agent to directly output strict JSON, which could compromise its conversational naturalness and reasoning flow, we employ a dedicated formatting model to handle this conversion, ensuring that the structured JSON is used strictly for objective judgment.

While the official TravelPlanner benchmark utilizes GPT-4o (OpenAI, 2024) as the default formatting model, the massive scale of trajectory generation required for RL training and evaluation renders proprietary API calls prohibitively expensive. Therefore, we adopt DeepSeek-V3.2-Exp (DeepSeek-AI, 2025) as our consistent formatting model. To address potential concerns regarding evaluation fairness, we emphasize two points: First, **all checkpoints and comparative models in our study are evaluated using this exact same formatting pipeline**, ensuring a strictly fair comparison. Second, prior to large-scale deployment, we conducted a rigorous offline validation on a subset of sampled plans. DeepSeek-V3.2-Exp demonstrated exceptional instruction-following capabilities for structured formatting, achieving an alignment rate with human expert annotations that is comparable, if not marginally superior, to GPT-

4o, particularly in avoiding minor information loss during complex natural-language-to-JSON conversions. This combination of cost-efficiency and high reliability makes it the optimal choice for our extensive RL evaluation pipeline.

C Implementation Details

In this section, we detail our implementation, covering data synthesis, the default experimental setups for SFT and RL, and the inference protocol. Specific configurations for individual ablation studies are deferred to their respective analysis sections.

C.1 Data Synthesis

During the element sampling process, we balance the ratio of selecting easy, medium, and hard element combinations. For query generation, we randomly select with equal probability from strong open-weight models: GPT-OSS-120B (OpenAI, 2025a) and DeepSeek-V3.2-Exp-671B (DeepSeek-AI, 2025). We utilize different models for query generation to ensure a more diverse expression of travel intents. For the sampled elements, we first query the sandbox to verify whether the specific dates and constraints yield feasible solutions. Only for the validated elements do we feed into the LLM to translate them into a natural language query. For example, given the input: { "org": "Everett", "dest": "San Francisco", "days": 3, "date": ["2022-06-16", "2022-06-17", "2022-06-18"], "people_number": 1, "hard_constraint": None, "budget": 1600 }, the DeepSeek-V3.2-Exp model translates it into the following query: "Could you help me plan a 3-day trip from Everett to San Francisco, from June 16th to June 18th, 2022, with a budget of \$1,600 for one person?"

Through this iterative process, we curated over 10K queries with balanced difficulty levels for subsequent use. To ensure reliability, we randomly sampled 200 cases from this pool, on which the DeepSeek-V3.2-Exp-Thinking model achieved a success rate of 21.9%, closely mirroring its 21.1% performance on the 180-instance TravelPlanner validation set.

C.2 Training Hyperparameters

SFT: From the synthetic pool, we prompted DeepSeek-V3.2-Exp-Thinking on 5K cases, filtering for task success and strict format adherence. This yielded 1,198 valid trajectories with

Table 9: **Statistics of filtered trajectories on synthetic data.** "Avg. Tokens of Planning" denotes the average token cost of the final generated itinerary.

Type	# Data Entries	Avg. Tool Calls	Avg. Tokens	Avg. Tokens of Planning
Easy	627 (52.3%)	8.8	10.1K	1204.7
Medium	373 (31.1%)	8.4	9.8K	1203.1
Hard	198 (16.5%)	11.7	11.8K	1343.1
All	1,198	9.2	10.3K	1227.1

their statistics shown in Table 9. Using the LLaMA-Factory framework¹, we performed full-parameter fine-tuning with a training batch size of 32, a learning rate of 5×10^{-6} , and a linear warmup ratio of 0.1. We trained the 3B and 7B models for 4 epochs, while the 1.5B model was trained for 6 epochs to compensate for its higher initial entropy. **RL:** Our default RL optimization utilizes the GRPO algorithm (Shao et al., 2024). The default training set comprises 1K synthetic samples (excluding SFT data) with an easy:medium:hard ratio of 4:3:3 to balance difficulty. We apply a maximum context length of 30K, a tool call budget of 60, a training batch size of 32, a group size of 8, a learning rate of 2×10^{-6} for 5 epochs, and a sampling temperature of 1.0. For reward computation, we first invoke DeepSeek-V3.2-Exp to parse the generated text plans into structured JSONs, enabling automated sub-metric evaluations, e.g., common-sense and hard constraints. The **CURRICULUM** reward is then implemented as: the SUM reward for the first two epochs, the MACRO reward for the next two, and the binary SUCCESS reward for the final epoch, forming a gradual transition from dense to sparse. Model selection is based on the best performance on TravelPlanner’s 180-instance validation set. Training for 7B models is distributed across two 8×NVIDIA A100-80G nodes, while 1.5B and 3B models utilize a single node.

C.3 Evaluation and Inference Setup

Evaluation spans in-domain testing on the 1,000-instance TravelPlanner test set and OOD testing across 7 distinct knowledge-intensive QA benchmarks including Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (2Wiki) (Ho et al., 2020), Musique (Trivedi et al., 2022), and Bamboogle (Press et al., 2023).

For in-domain evaluation, inference on the TravelPlanner benchmark follows the standard ReAct

¹<https://github.com/hiyouga/LlamaFactory>

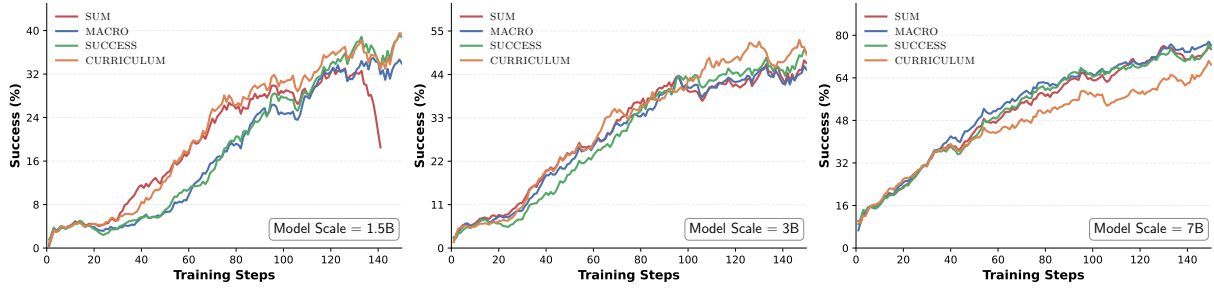


Figure 8: **Training dynamics (i.e., success rates on the training samples during each step) of different rewards.** The **CURRICULUM** reward accelerates convergence and achieves higher peak performance for smaller models. Conversely, the dense rewards prove most effective for the stronger 7B model.

paradigm. The agent performs iterative cycles of thinking, issuing tool calls, and receiving environmental feedback. The final natural language itinerary is parsed into a structured JSON format using DeepSeek-V3.2-Exp for automated evaluation via the official TravelPlanner Leaderboard². To ensure fair comparisons, **all models are evaluated using this identical pipeline and codebase.**

- **SOTA LLMs:** We evaluate recent strong models including DeepSeek-V3.2-Exp-Thinking-671B (DeepSeek-AI, 2025), Gemini3-Pro (Team, 2025), Seed-1.8 (Seed, 2026), Kimi-K2.5-1T (Team, 2026a), and Qwen3.5-series (Team, 2026b). We set the maximum context length to 128K and the tool call budget to 100. The temperature is set to 1.0 for Kimi-K2.5-1T and 0.6 for others. We also include GPT-5 (OpenAI, 2025b) and Planner-R1-32B results from Zhu et al. (2025).
- **Trained Checkpoints:** For our trained models and their Qwen2.5-Instruct base models, we consistently apply a temperature of 0.6, a top_p of 0.95, a maximum context length of 32K, and a tool call budget of 60.

For OOD evaluation, we utilize each dataset’s official test split. The only available tool is a Search function that queries a local Wikipedia corpus. Following the practice of Search-R1 (Jin et al., 2025b), we employ E5 (Wang et al., 2024) as the retriever. For each query, we retrieve the top-5 text snippets. Additionally, the maximum context length is set to 32K, with a tool call budget of 60. For all inference, trajectories exceeding the context limit, tool calling budget, or failing to output the required `<answer>`

²<https://huggingface.co/spaces/osunlp/TravelPlannerLeaderboard>

`</answer>` format are terminated and marked as failures.

Deep-Research OOD Evaluation: To probe transfer beyond knowledge-intensive QA, we additionally evaluate on BrowseComp-Plus (Chen et al., 2025), a deep-research benchmark built over a fixed, human-verified corpus so that agent capability can be disentangled from retriever quality. We use Qwen3-Embedding-8B as the retriever with top- $K = 5$, keeping the same Search-only action space, 32K context, and tool call budget of 60 as above. We report answer accuracy, evidence recall, i.e., the fraction of human-annotated supporting documents surfaced by the agent, and the average number of search actions per query, which serves as a proxy for search efficiency. For reference, we additionally evaluate Search-R1-32B (Jin et al., 2025b) under the same retriever setup.

D Supplementary Materials for Experiments

D.1 Reward Shaping

Setup: To ensure a fair comparison, the only variable modified in this ablation is the reward computation strategy. All other settings strictly adhere to the default hyperparameter configuration. The specific epoch-level transitions for the **CURRICULUM** strategy have been detailed in the previous implementation section.

Training Dynamics: Figure 8 illustrates the success rate curves over training steps, providing granular evidence for **Takeaway 1**, i.e., scale-dependent reward design. For the 1.5B and 3B models, the **CURRICULUM** reward maintains the upper bound of performance, accelerating convergence. This shows that smaller models require progressive, staged learning signals. In contrast, the 7B model exhibits robust learning capacity, allowing

Table 10: **Out-of-domain performance (in %) on the knowledge-intensive QA benchmarks.** Results marked with * are sourced from Ji et al. (2026). The best results of each dataset across methods are highlighted in **bold**. Our RL-trained models achieve comparable or superior performance to the domain-specific Search-R1 baseline (Jin et al., 2025b) across model scales, showing that complex tool-calling capabilities effectively generalize to other domains.

Scale	Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
1.5B	Base*	7.1	22.4	9.9	5.9	4.3	2.6	8.0	8.6
	SFT	13.5	21.4	13.5	11.6	7.0	1.2	12.0	11.5
	Search-R1*	39.4	51.0	39.7	14.6	24.4	2.2	4.0	25.0
	SUM	32.1	44.5	30.9	26.0	14.8	5.3	16.8	24.3
	MACRO	29.6	45.1	28.8	26.9	19.4	6.3	23.2	25.6
	SUCCESS	24.8	36.9	21.3	24.2	17.7	4.6	23.2	21.8
	CURRICULUM	20.6	28.5	20.5	16.0	11.7	3.5	17.6	16.9
3B	Base	10.6	28.8	10.8	14.9	24.4	2.0	2.4	13.4
	SFT	35.1	52.5	31.3	32.0	24.4	9.0	30.4	30.7
	Search-R1	34.1	54.5	37.8	32.4	31.9	10.3	26.4	32.5
	SUM	38.2	54.9	34.3	34.2	23.3	8.8	27.2	31.6
	MACRO	37.0	54.5	34.9	37.9	29.9	11.6	33.6	34.2
	SUCCESS	37.5	54.1	34.0	33.7	19.4	10.3	32.0	31.6
	CURRICULUM	41.0	56.8	36.2	39.5	27.7	12.4	32.0	35.0
7B	Base	13.4	40.8	14.0	18.3	25.0	3.1	12.0	18.1
	SFT	41.2	59.7	37.6	47.0	37.8	16.3	53.6	41.9
	Search-R1	39.3	61.0	39.7	37.0	41.4	14.6	36.8	38.5
	SUM	35.5	54.5	34.8	44.8	34.6	15.4	37.6	36.7
	MACRO	42.2	61.2	39.6	48.8	38.3	17.4	52.8	42.9
	SUCCESS	38.7	55.9	35.2	45.8	38.4	15.5	41.6	38.7
	CURRICULUM	41.1	58.7	37.7	48.4	38.5	17.4	45.6	41.1

the dense rewards to efficiently and stably optimize performance without requiring curriculum guidance.

OOD Generalization: Table 10 details the OOD performance of different reward strategies on 7 knowledge-intensive QA benchmarks. To contextualize our results, we include three baselines: direct inference of the Base model, the SFT model, and Search-R1 (Jin et al., 2025b), a strong baseline explicitly trained on its respective domain data. The results offer two critical insights: (1) **Transferability of complex tool-use:** Models trained exclusively on our TravelPlanner trajectories develop reasoning capabilities that transfer remarkably well to simpler domains. Across multiple scales, our RL variants achieve comparable or superior performance to the domain-specific Search-R1 baseline. (2) **Evidence of the alignment tax:** Table 10 provides the empirical foundation for **Takeaway 2**. For the 7B model, while the dense SUM reward maximizes in-domain success (Table 1), its OOD average accuracy drops to 36.7%, falling significantly behind its SFT starting point (41.9%).

This demonstrates that overly dense, task-specific rewards cause the model to overfit to the TravelPlanner format, degrading general information-seeking abilities. The semi-sparse MACRO reward mitigates this alignment tax, achieving the highest OOD performance (42.9%) while remaining highly competitive in-domain.

D.2 Model Scaling

Training Dynamics: To provide a more granular view of the scaling behaviors discussed in **Takeaway 3**, we present the training dynamics of the 1.5B, 3B, and 7B models across all four reward configurations in Figure 9. Scaling up the base model capacity yields three distinct advantages during the RL fine-tuning process: **accelerated convergence**, **enhanced training stability**, and **higher performance asymptotes**. Across all reward designs, the 7B models exhibit a significantly steeper initial learning curve compared to their 1.5B and 3B counterparts. For instance, under the sparse SUCCESS reward, the 1.5B model struggles to establish a clear upward sign in the early stages, whereas the

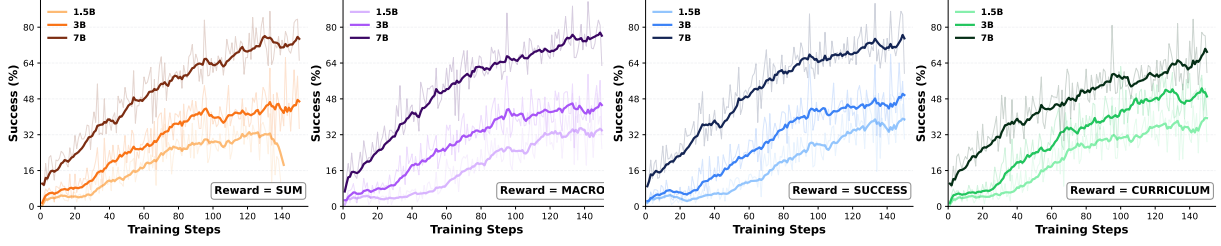


Figure 9: **Training dynamics across model scales.** For all reward types, increasing model capacity from 1.5B to 7B consistently leads to faster convergence, improved training stability, and higher performance asymptotes.

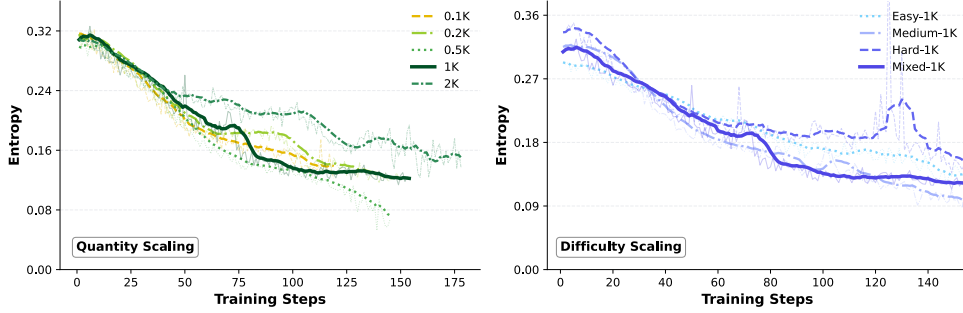


Figure 10: **Training dynamics for data quantity and difficulty experiments.** Configurations that yield the best generalization performance (1K and Mixed) are characterized by a stable, gradual decline in entropy. Suboptimal configurations suffer from anomalous training behaviors, e.g., premature entropy collapse, severe training instability.

7B model rapidly acquires the necessary reasoning patterns, indicating that larger models possess stronger inherent priors to leverage RL signals efficiently. Beyond learning faster, larger models demonstrate superior robustness against the high variance typically associated with RL. The learning curves for the 1.5B model, exhibit fluctuations, whereas the 7B model maintains smoothness. Ultimately, larger models converge to higher final success rates, with the performance gaps widening over the course of training. This shows that the bottlenecks observed in smaller models are fundamental capacity limits.

D.3 Data Composition

Data Construction: For all data composition experiments, we utilize the 3B model and apply the CURRICULUM reward to isolate the variables of data quantity and difficulty.

- **Quantity:** The baseline is the 1K prompt set, which follows a 4:3:3 ratio of easy, medium, and hard samples. To create the 100, 200, and 500 subsets, we randomly subsample from this 1K baseline while preserving the 4:3:3 difficulty distribution. For the 2K setting, we augment the baseline with an additional 1K samples from the synthetic pool, adhering to the same difficulty ratio.

- **Difficulty:** We independently construct three datasets, Easy-1K, Medium-1K, and Hard-1K, by sampling 1K instances comprised entirely of their respective difficulty levels from the remaining synthetic pool. These are compared against the default Mixed-1K baseline.

Setup: Varying the dataset size introduces a critical confounding factor: the number of gradient updates per epoch. To ensure a fair comparison and isolate the effect of the data itself, we adjust the total training epochs for different subset sizes. This guarantees that all models undergo **a total of approximately 155 gradient update steps**, which is equivalent to 5 epochs on the 1K dataset with a batch size of 32.

Training Dynamics & OOD Results: Figure 10 illustrates the training dynamics through action entropy, which reflects the model’s exploration behavior. For data quantity, the 1K configuration exhibits a smooth, gradual decline in entropy, whereas settings like 0.5K display premature entropy collapse, indicating potential over-optimization. Regarding data difficulty, the Mixed-1K setting maintains a stable entropy descent. In contrast, the Hard-1K configuration suffers from severe training instability. Finally, **Table 11** provides the comprehensive OOD results. Consistent with our in-domain findings, the 1K dataset size and the Mixed difficulty

Table 11: **Out-of-domain performance (in %) on the knowledge-intensive QA benchmarks.** The best results of each dataset across data configurations are highlighted in **bold**.

Mode	Data	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
	SFT	35.1	52.5	31.3	32.0	24.4	9.0	30.4	30.7
<i>Quantity</i>	0.1K	39.2	57.4	35.0	38.5	26.3	9.8	30.4	33.8
	0.2K	37.5	56.7	34.7	36.8	29.2	9.6	29.6	33.4
	0.5K	39.9	57.3	34.0	37.4	25.2	11.8	29.6	33.6
	1K	41.0	56.8	36.2	39.5	27.7	12.4	32.0	35.0
	2K	38.5	56.7	34.4	34.3	22.4	9.6	29.6	32.2
<i>Difficulty</i>	Easy-1K	37.0	54.4	33.5	37.7	30.7	10.1	31.4	33.5
	Medium-1K	38.7	56.5	34.9	38.3	26.5	10.7	26.4	33.1
	Hard-1K	38.1	55.1	36.8	39.0	29.9	11.1	28.0	34.0
	Mixed-1K	41.0	56.8	36.2	39.5	27.7	12.4	32.0	35.0



Figure 11: **Training dynamics of different RL algorithms.** The plots display the success rate curves during training for **1.5B (left)**, **3B (middle)**, and **7B (right)** models. Exploration-heavy algorithms like ARPO exhibit clear advantages at smaller scales, but this benefit diminishes as model capacity increases.

composition achieve the highest average OOD accuracy in this experiment.

D.4 Algorithm Selection

Setup: To isolate the RL algorithm as the sole variable, all configurations utilize the default training data and hyperparameters. For the 1.5B and 3B models, although the CURRICULUM reward empirically yields the best performance, we apply the MACRO reward to eliminate potential confounding factors introduced by reward switching. For the 7B model, we apply its optimal SUM reward. Across all three algorithms, technical enhancements designed to stabilize training, e.g., KL-free loss, remain consistent. Additionally, for DAPO, we maintain a batch size of 32 without further over-sampling.

Training Dynamics: Figure 11 illustrates the training success rate curves for different RL algorithms. These dynamics further corroborate our findings in the main text: for smaller models, exploration-heavy algorithms demonstrate a clear advantage over standard GRPO. However, as model capacity scales to 3B and 7B, this algorithmic advantage diminishes. GRPO smoothly matches or even surpasses the complex variants, visually confirming

that stronger base models rely less on heuristic exploration mechanisms.

D.5 Environmental Stability

We study environmental instability by randomly injecting perturbations only during RL training, while evaluating all resulting checkpoints in the same clean, failure-free TravelPlanner test environment. We consider three perturbation types.

- **API Dropout:** During tool execution, the environment returns a global error message `<tool_response> Error: Current tool {tool_name} is not available. </tool_response>` with a predefined probability.
- **Missing-field:** For information-retrieval tools, one randomly selected attribute in each returned record is masked as `[MISSING]` with probability p .
- **Missing + Distractor:** In addition to missing fields, we inject irrelevant records into the returned observations, e.g., restaurants from other cities when searching for the target city.

For the observation-corruption experiments, we use the 3B model with the MACRO reward and set $p \in \{0.05, 0.10\}$. Across all settings, we keep the training data and other hyperparameters unchanged.

D.6 Recipe Summary

Building on the preceding analyses, we synthesize the optimal training recipes tailored to each model scale. Regarding data configuration, we consistently apply the Mixed-1K dataset across all models, following the configuration identified in our primary data-composition study. For the 1.5B model, we adopt the ARPO algorithm to encourage exploration under high uncertainty. For the 3B model, we transition to the GRPO algorithm with the CURRICULUM reward. Finally, for the 7B model, we utilize GRPO combined with the dense SUM reward. The models trained with these specific configurations correspond to the checkpoints evaluated against leading LLMs in **Figure 1**.

E Cases

In this section, we present the full trajectories on TravelPlanner’s validation set generated by our STAR-trained agents, specifically focusing on the best checkpoints across different model scales.

We first showcase successful cases where the agent intelligently uses tools to gather relevant information, effectively balancing available choices for accommodations, meals, and transportation. This allows the agent to control the budget while strictly satisfying the user’s requirements. Notably, in the first case, when the agent discovers that ground transportation is infeasible, it autonomously pivots to searching for flights. This demonstrates an emergent ability for *local error correction* and persistent information gathering. In the second case, the agent exhibits robust long-horizon planning, seamlessly managing a multi-city itinerary under complex group and accommodation constraints.

To provide a comprehensive analysis and highlight avenues for future research, we also present failure cases that expose the inherent limitations of current LLM-based agents. First, we show a case illustrating that the agent can still struggle with *dense constraints*. For example, the agent fails to fulfill the user’s highly specific request for 4 distinct types of cuisines, indicating a degradation in instruction following when cognitive load is high.

More importantly, we highlight a critical struc-

tural limitation in complex planning: *the lack of global backtracking and re-planning*. In our final failure case, the agent selects a destination city but subsequently discovers through tool calls that no accommodations in that city meet the user’s strict requirements, e.g., a “private room” that allows “parties”. Instead of recognizing this as a dead end in the state space and backtracking to select a different city, the agent stubbornly proceeds. It compromises the hard constraints by selecting an invalid room type to justify its flawed trajectory.

Future Direction: This contrasting behavior, succeeding in local adaptation but failing in global re-planning, underscores a crucial gap in current agentic capabilities. It suggests that while our method significantly improves sequential decision-making, LLMs still struggle with long-horizon dependencies where upstream decisions must be revised based on downstream observations. Addressing this bottleneck remains a highly promising direction for future work in agent planning and reasoning.

F Prompt

We further detail the prompts used in our evaluation. These include the system prompt for TravelPlanner inference, the system prompt for knowledge-intensive QA tasks inference, and the prompt used to convert natural language plans into JSON-structured formats for judgment.

Successful trajectory on TravelPlanner

Question: Please devise a travel plan that starts from **St. Petersburg** and heads to **Appleton**, taking place across **3 days** from **March 19th to March 21st, 2022**. This itinerary is for an individual, with a budget allocated at **\$1,200**.

Turns 1 - 5: Transportation Planning & Error Handling

<think>

I need to start by checking the travel time and cost from St. Petersburg to Appleton on March 19th. I'll use GoogleDistanceMatrix to get the details.

</think>

<tool_call>

{ "name": "GoogleDistanceMatrix", "arguments": { "departure": "St. Petersburg", "destination": "Appleton", "mode": "driving" } }

</tool_call>

<tool_response>

Sorry, we cannot find the distance between St. Petersburg and Appleton.

</tool_response>

<think>

Since there's no direct route from St. Petersburg to Appleton via GoogleDistanceMatrix, I need to try another city in Wisconsin that might have better connectivity. I'll check Milwaukee first.

</think>

<tool_call>

{ "name": "GoogleDistanceMatrix", "arguments": { "departure": "St. Petersburg", "destination": "Milwaukee", "mode": "driving" } }

</tool_call>

<tool_response>

No valid information found. Please rethink your distance search and try again.

</tool_response>

<think>

Since there's no direct route to Appleton or Milwaukee via GoogleDistanceMatrix, I need to try a different city in Wisconsin. Green Bay might have better connectivity.

</think>

<tool_call>

{ "name": "GoogleDistanceMatrix", "arguments": { "departure": "St. Petersburg", "destination": "Green Bay", "mode": "driving" } }

</tool_call>

<tool_response>

No valid information found. Please rethink your distance search and try again.

</tool_response>

<think>

Since there's no direct route to Appleton or nearby cities via GoogleDistanceMatrix, I need to try a different approach. I'll search for flights from St. Petersburg to Appleton on March 19th.

</think>

<tool_call>

{ "name": "SearchFlight", "arguments": { "departure": "St. Petersburg", "destination": "Appleton", "date": "2022-03-19" } }

</tool_call>

<tool_response>

A SearchFlight for St. Petersburg to Appleton on 2022-03-19 found the following 1 flights:

1. [F3574992] \$542. Departure: 22:51, Arrival: 00:42, Actual Elapsed Time: 2 hours 51 minutes, Distance: 1173.0 miles

</tool_response>

<think>

Now I need to search for the return flight from Appleton to St. Petersburg on March 21st.

</think>

<tool_call>

{ "name": "SearchFlight", "arguments": { "departure": "Appleton", "destination": "St. Petersburg", "date": "2022-03-21" } }

</tool_call>

<tool_response>

A SearchFlight for Appleton to St. Petersburg on 2022-03-21 found the following 1 flights:

1. [F3578689] \$332. Departure: 07:19, Arrival: 11:01, Actual Elapsed Time: 2 hours 42 minutes, Distance: 1173.0 miles

</tool_response>

Turns 6 - 8: Accommodation, Attractions & Dining Search

<think>

Now I need to search for accommodation in Appleton to complete the travel plan.

```

</think>
<tool_call>
{"name": "SearchAccommodation", "arguments": {"city": "Appleton"}}
</tool_call>
<tool_response>
A SearchAccommodation for Appleton found the following 14 accommodations:
1. [Nice bedroom in Brooklyn !] $932.0. Room Type: Private room, House Rules: No pets, Minimum Nights: 4.0, Maximum Occupancy: 1, Review Rate Number: 5.0
2. [Cozy room + bright apartment + rooftop lounge] $666.0. Room Type: Private room, House Rules: No smoking, Minimum Nights: 1.0, Maximum Occupancy: 1, Review Rate Number: 4.0
...
</tool_response>
<think>
Now I need to search for attractions in Appleton to continue building the itinerary.
</think>
<tool_call>
{"name": "SearchAttraction", "arguments": {"city": "Appleton"}}
</tool_call>
<tool_response>
A SearchAttraction for city Appleton found the following 20 attractions:
1. [The History Museum at the Castle](http://www.myhistorymuseum.org/) Address: 330 E College Ave, Appleton, WI 54911, USA. Contact: (920) 735-9370
2. [Hearthstone Historic House Museum](https://www.hearthstonemuseum.org/) Address: 625 W Prospect Ave, Appleton, WI 54911, USA. Contact: (920) 730-8204
...
</tool_response>
<think>
Now I need to search for restaurants in Appleton to complete the travel plan.
</think>
<tool_call>
{"name": "SearchRestaurant", "arguments": {"city": "Appleton"}}
</tool_call>
<tool_response>
A SearchRestaurant for city Appleton found the following 36 restaurants:
1. [Mikata Japanese Steakhouse] Avg. Cost $72. Cuisines: Cafe, Pizza, Bakery, Italian. Rating: 3.6
2. [Brick 29] Avg. Cost $97. Cuisines: Desserts, Tea, BBQ, Cafe, American, Mediterranean. Rating: 4.4
...
</tool_response>

```

Turn 9: Final Itinerary Synthesis

```

<think>
I have gathered all the necessary information from the tool calls. I need to compile the final travel plan with a day-by-day itinerary.
</think>
<answer>

```

Travel Plan: St. Petersburg to Appleton (March 19-21, 2022)

Budget: \$1,200 | Traveler: Individual

Day 1: March 19 - Travel to Appleton

- Morning/Afternoon: **Flight** from **St. Petersburg** to **Appleton** [F3574992] | **\$542** | Departure: 22:51 | Arrival: 00:42
- Evening: Check into accommodation - [**Stylish, convenient, renovated- 2 min to subway**] | **\$64/night** | Minimum Nights: 1 | Entire home/apt

Day 2: March 20 - Explore Appleton

- Breakfast: At **New Bakers Shoppee** | Avg Cost: **\$15**
- Morning Attraction: Visit **Atlas Science Center**
- Lunch: At **Side Wok** | Avg Cost: **\$15**
- Afternoon Attraction: Visit **Trout Museum of Art**
- Dinner: At **Mathew's Cafe** | Avg Cost: **\$13**

Day 3: March 21 - Return Travel

- Morning: **Flight** from **Appleton** to **St. Petersburg** [F3578689] | **\$332** | Departure: 07:19 | Arrival: 11:01

Key Highlights

- **Total Budget: \$1,200**

- **Accommodation:** **Stylish, convenient** (2 nights) | **\$128** total
- **Flights:** **\$874** total

</answer> Satisfy both commonsense and hard constraints to achieve *Success* ✓

Successful trajectory on TravelPlanner

Question: Could you help develop a **week-long** travel itinerary suitable for a group of 6 people, departing from **Baton Rouge** and planning to visit **3 different cities in Texas**? The travel dates are set from **March 17th to March 23rd, 2022**. Our travel budget has been adjusted to **\$14,600**. Bearing in mind that we have **children under ten years old**, our accommodations need to allow young children and we prefer to occupy **entire rooms**. We also prefer **not to self-drive** during this trip.

Turns 1 - 5: Destination Selection & Transportation Planning

<think>

I need to start by **finding cities in Texas** to select three for the itinerary. I'll use the SearchCity tool to get a list of cities in Texas.

</think>

<tool_call>

{ "name": "SearchCity", "arguments": { "state": "Texas" } }

</tool_call>

<tool_response>

A SearchCity for state Texas found the following 23 cities: Abilene, Amarillo, Harlingen, Lubbock, College Station, Corpus Christi, Wichita Falls, Waco, San Angelo, Houston, San Antonio, ...

</tool_response>

<think>

I need to start by checking travel options from Baton Rouge to the first city. I'll **choose Houston as the first city** since it's a major city with good connectivity. I'll use GoogleDistanceMatrix to get travel details from Baton Rouge to Houston.

</think>

<tool_call>

{ "name": "GoogleDistanceMatrix", "arguments": { "departure": "Baton Rouge", "destination": "Houston", "mode": "taxi" } }

</tool_call>

<tool_response>

A GoogleDistanceMatrix found the following information: taxi, from Baton Rouge to Houston, duration: 4 hours 0 mins, distance: 431 km, cost: \$431

</tool_response>

<think>

Now I need to check travel from Houston to the second city. I'll **choose Austin as the second city** since it's a key city in Texas. I'll use GoogleDistanceMatrix to get travel details from Houston to Austin.

</think>

<tool_call>

{ "name": "GoogleDistanceMatrix", "arguments": { "departure": "Houston", "destination": "Austin", "mode": "taxi" } }

</tool_call>

<tool_response>

A GoogleDistanceMatrix found the following information: taxi, from Houston to Austin, duration: 2 hours 31 mins, distance: 266 km, cost: \$266

</tool_response>

... [Model successfully repeats GoogleDistanceMatrix calls for Austin → San Antonio, and San Antonio → Baton Rouge using Taxi]

Turns 6 - 8: Accommodation Search

<think>

Now I need to search for accommodations in Houston. I'll use SearchAccommodation.

</think>

<tool_call>

{ "name": "SearchAccommodation", "arguments": { "city": "Houston" } }

</tool_call>

<tool_response>

A SearchAccommodation for Houston found the following 10 accommodations:

1. [Sunny and Airy near Manhattan] \$499.0. Room Type: Entire home/apt, House Rules: No pets, Maximum Occupancy: 4...
2. [Bright room in 2BR apt in Nolita!] \$1024.0. Room Type: Private room, House Rules: No children under 10...
3. [Superhost 3 bedroom DISCOUNT] \$746.0. Room Type: Entire home/apt, House Rules: No pets, Minimum Nights: 1.0, Maximum Occupancy: ...

...

</tool_response>

... [Model successfully repeats SearchAccommodation calls for Austin and San Antonio]

Turns 9 - 14: Attractions & Dining Search

<think>

Now I need to search for attractions in Houston to start building the itinerary.

</think>

<tool_call>

{“name”: “SearchAttraction”, “arguments”: {“city”: “Houston”}}

</tool_call>

<tool_response>

A SearchAttraction for city Houston found the following 20 attractions:

1. [Downtown Aquarium](http://www.aquariumrestaurants.com/downtownaquariumhouston/) Address: 410 Bagby St, Houston, TX 77002, USA. Contact: (713) 223-3474
2. [Space Center Houston](https://www.spacecenter.org/) Address: 1601 E NASA Pkwy, Houston, TX 77058, USA. Contact: (281) 244-2100
3. [Water Wall](https://uptown-houston.com/experience/parks/waterwall-park/) Address: 2800 Post Oak Blvd, Houston, TX 77056, USA. Contact: (713) 850-8841

...

</tool_response>

... [Model continues to search for attractions in Austin and San Antonio, and then repeats SearchRestaurant for the three cities]

Turn 15: Final Itinerary Synthesis

<think>

I have gathered all the necessary information from the tool calls. I need to synthesize this into a comprehensive travel itinerary for a group of 6 people with a budget of \$14,600, visiting Houston, Austin, and San Antonio from March 17th to March 23rd, 2022. The itinerary will include travel details, accommodations, attractions, and meals.

</think>

<answer>

Week-long Travel Itinerary for 6 People

Budget: \$14,600 | **Dates: March 17–23, 2022** | **No Self-Driving** | **Children Under 10** | **Entire Room Preferences**

Day 1: March 17 – Travel to Houston

- Morning: **Taxi** from **Baton Rouge** to **Houston** | Cost: **\$431** | Duration: 4h
- Afternoon: Check into [**Superhost 3 bedroom DISCOUNT**] | Entire home/apt | Cost: **\$746/night**
- Dinner: At **Vinayaka Mylari** | Avg Cost: **\$14/person** | Cuisines: Fast Food, French, Cafe, Italian
- Attraction: Visit **Hermann Park** | Contact: (713) 526-2183

Day 2: March 18 – Explore Houston

- Breakfast: At **Super Bakery** | Avg Cost: **\$30/person** | Cuisines: Seafood, Fast Food
- Morning Attraction: **Houston Museum of Natural Science** | Contact: (713) 639-4629
- Lunch: At **Matchbox** | Avg Cost: **\$33/person** | Cuisines: Tea, Italian, Bakery, Fast Food, Seafood
- Afternoon Attraction: **Space Center Houston** | Contact: (281) 244-2100
- Dinner: At **Earthen Spices** | Avg Cost: **\$28/person** | Cuisines: Tea, Mexican, Bakery, BBQ, Fast Food

Day 3: March 19 – Travel to Austin

- Morning: **Taxi** from **Houston** to **Austin** | Cost: **\$266** | Duration: 2h 31m
- Afternoon: Check into [**Spacious Williamsburg 1 bedroom**] | Entire home/apt | Cost: **\$455/night**
- Dinner: At **Moksha** | Avg Cost: **\$12/person** | Cuisines: Cafe, Bakery, BBQ, Fast Food
- Attraction: Visit **Zilker Metropolitan Park** | Contact: (512) 974-6700

... [Days 4-6 omitted for brevity, successfully detailing **Austin** and **San Antonio** with **Taxi** transfers, **Entire home/apt** stays, and 3 **meals** per day]

Day 7: March 23 – Return to Baton Rouge

- Morning: **Taxi** from **San Antonio** to **Baton Rouge** | Cost: **\$746** | Duration: 6h 52m

Budget Summary

- Total Budget: **\$14,600**
- Estimated Costs: **\$13,500** (within budget)
- Remaining: **\$1,100**

Key Highlights

- Accommodations: **Entire rooms selected for children.**
- Transport: **Taxi** used for all segments
- Attractions: Visited key sites in each city
- Meals: Varied **cuisines** covered

</answer> **Satisfy both commonsense and hard constraints to achieve Success ✓**

Failure trajectory on TravelPlanner: Incomplete Constraint Satisfaction

Question: Can you please generate a 3-day travel plan for 2 people departing from **New York** and traveling to **Charleston** from **March 19th to March 21st, 2022**? Our budget is set at **\$1,600**. We require accommodations that **allow smoking** and should **not be shared rooms**. We have a preference for **French, Mediterranean, Mexican, and Chinese cuisines** during our trip.

Turns 1 - 2: Transportation Planning

<think>

I need to find flight information from New York to Charleston on March 19th, 2022, as this is the departure date for the travel plan.

</think>

<tool_call>

{“name”: “SearchFlight”, “arguments”: {“departure”: “New York”, “destination”: “Charleston”, “date”: “2022-03-19”}}

</tool_call>

<tool_response>

A SearchFlight for New York to Charleston on 2022-03-19 found the following 8 flights:

...

8. [F4066693] \$137. Departure: 17:48, Arrival: 19:56, Actual Elapsed Time: 2 hours 8 minutes, Distance: 641.0 miles

</tool_response>

<think>

I need to find return flight information from Charleston to New York on March 21st, 2022, as this is the return date for the travel plan.

</think>

<tool_call>

{“name”: “SearchFlight”, “arguments”: {“departure”: “Charleston”, “destination”: “New York”, “date”: “2022-03-21”}}

</tool_call>

<tool_response>

A SearchFlight for Charleston to New York on 2022-03-21 found the following 11 flights:

1. [F3635917] \$143. Departure: 11:01, Arrival: 12:58, Actual Elapsed Time: 1 hours 57 minutes, Distance: 636.0 miles

...

</tool_response>

Turn 3: Accommodation Search

<think>

I need to find accommodation options in Charleston that allow smoking and are not shared rooms, as per the user’s requirements.

</think>

<tool_call>

{“name”: “SearchAccommodation”, “arguments”: {“city”: “Charleston”}}

</tool_call>

<tool_response>

A SearchAccommodation for Charleston found the following 20 accommodations:

1. [2 Bedroom Apartment in Prime Midtown West] \$1121.0. Room Type: Entire home/apt, House Rules: No smoking & No children under 10...

2. [Bronx Room] \$127.0. Room Type: Private room, House Rules: No visitors & No children under 10, Minimum Nights: 2.0, Maximum Occupancy: 2, Review Rate Number: 2.0

...

</tool_response>

Turns 4 - 5: Attractions & Dining Search

<think>

I need to find restaurant options in Charleston that match the user’s preferred cuisines: French, Mediterranean, Mexican, and Chinese.

</think>

<tool_call>

{“name”: “SearchRestaurant”, “arguments”: {“city”: “Charleston”}}

</tool_call>

<tool_response>

A SearchRestaurant for city Charleston found the following 62 restaurants:

1. [Filling Station] Avg. Cost \$78. Cuisines: Desserts, Tea, Italian, **French**, Cafe. Rating: 4.1

...

4. [Makansutra Gluttons Bay] Avg. Cost \$84. Cuisines: Tea, **French**, BBQ, **Mediterranean**. Rating: 3.0

...

18. [The Culinary Pitaara] Avg. Cost \$23. Cuisines: Tea, Italian, Bakery, BBQ, Cafe. Rating: 4.2

...

60. [Roka] Avg. Cost \$84. Cuisines: Desserts, Tea, **Mexican**, Fast Food, Cafe, **Mediterranean**. Rating: 4.6

...

</tool_response>

<think>

I need to find tourist attractions in Charleston to include in the daily itinerary for the 3-day trip.

</think>

<tool_call>

{“name”: “SearchAttraction”, “arguments”: {“city”: “Charleston”}}

</tool_call>

<tool_response>

A SearchAttraction for city Charleston found the following 20 attractions:

1. [South Carolina Aquarium](http://scaquarium.org/) Address: 100 Aquarium Wharf, Charleston, SC 29401, USA. Contact: (843) 577-3474

...

</tool_response>

Turn 6: Final Itinerary Synthesis

<think>

Based on the gathered information, I will create a comprehensive 3-day travel plan from New York to Charleston from March 19th to March 21st, 2022, for 2 people with a budget of \$1,600. The plan includes:

- Flights: Selected the most cost-effective options for both directions.
- Accommodation: Chose “Bronx Roon” which allows smoking and is not shared, with minimum nights of 2 as required.
- Restaurants: Selected French, Mediterranean, Mexican, and Chinese cuisines as per preference, with variety in timing and cost.
- Attractions: Included popular Charleston attractions to cover the itinerary.
- Budget: Total estimated cost is within the \$1,600 budget.

The plan is day-by-day, ensuring all requirements are met.

</think>

<answer>

3-Day Charleston Travel Plan from New York

Dates: March 19-21, 2022 | Travelers: 2 people | Total Budget: **\$1,600**

Day 1: March 19 (Travel from New York to Charleston)

- Morning: Depart from **New York** on **Flight** [F4066693] at 17:48, arriving in **Charleston** at 19:56. Cost: **\$137** per person $\times 2 =$ **\$274**
- Evening: Check into accommodation: [**Bronx Roon**] (Private room, House Rules: No visitors, Maximum Occupancy: 2). Cost: **\$127** per night $\times 2$ nights = **\$254**
- Dinner: **Makansutra Gluttons Bay** (French, Mediterranean). Avg. Cost: **\$84** per person $\times 2 =$ **\$168**. Rating: 3.0.

Day 2: March 20 (Full day in Charleston)

- Breakfast: **Filling Station** (French, Cafe). Avg. Cost: **\$78** per person $\times 2 =$ **\$156**. Rating: 4.1.
- Lunch: **The Culinary Pitaara** (Italian, Bakery). Avg. Cost: **\$23** per person $\times 2 =$ **\$46**. Rating: 4.2.
- Afternoon: Visit **South Carolina Aquarium** (Educational, Outdoor). Address: 100 Aquarium Wharf. Contact: (843) 577-3474.
- Dinner: **Roka** (Mexican, Mediterranean). Avg. Cost: **\$84** per person $\times 2 =$ **\$168**. Rating: 4.6.

Day 3: March 21 (Travel from Charleston to New York)

- Morning: Check out from [**Bronx Roon**] by 11:00 AM.
- Flight: Depart from **Charleston** on **Flight** [F3635917] at 11:01, arriving in **New York** at 12:58. Cost: **\$143** per person $\times 2 =$ **\$286**.

Cost Breakdown:

- Flights: **\$274** (to Charleston) + **\$286** (return) = **\$560**
- Accommodation: **\$254** for 2 nights
- Meals: **\$168** (Dinner Day 1) + **\$156** (Breakfast Day 2) + **\$46** (Lunch Day 2) + **\$168** (Dinner Day 2) = **\$538**
- Total: **\$560** + **\$254** + **\$538** = **\$1,352**

Notes:

- All accommodations and restaurants are selected based on your preferences.
- Attractions cover a mix of cultural, historical, and recreational activities.
- **Smoking is allowed** at the accommodation, and it's a **private room**.
- Budget has room for additional expenses or upgrades.

</answer>

Failed to satisfy hard constraints. The model hallucinated satisfying all cuisines but actually missing the required Chinese cuisine. ×

Failure trajectory on TravelPlanner: Lack of Global Backtracking

Question: Can you assist in crafting a travel itinerary for a pair of travelers originating from **Milwaukee** and visiting **2 cities in Michigan**? The trip is **5 days** long, spanning from **March 21st to March 25th, 2022**. Our budget is now set at **\$1,700**, and we require accommodations that **allow parties** and provide **private rooms**. Moreover, we'd prefer if our transportation **did not involve any flights**.

Turns 1 - 4: Destination Selection & Transportation Planning

<think>

I need to **find cities in Michigan** to select two for the itinerary. I'll start by searching for cities in Michigan.

</think>

<tool_call>

{"name": "SearchCity", "arguments": {"state": "Michigan"}}

</tool_call>

<tool_response>

A SearchCity for state Michigan found the following 12 cities: Pellston, Traverse City, Alpena, Iron Mountain, Kalamazoo, Saginaw, Grand Rapids, Lansing, Muskegon, Hancock, Detroit, Escanaba

</tool_response>

<think>

I need to check the travel time and cost from Milwaukee to the first city. I'll **start with Detroit** as it's a major city with good connectivity.

</think>

<tool_call>

{"name": "GoogleDistanceMatrix", "arguments": {"departure": "Milwaukee", "destination": "Detroit", "mode": "driving"}}

</tool_call>

<tool_response>

A GoogleDistanceMatrix found the following information: driving, from Milwaukee to Detroit, duration: 5 hours 40 mins, distance: 614 km, cost: \$30

</tool_response>

... [Model successfully plans driving routes: Detroit → Grand Rapids, and Grand Rapids → Milwaukee]

Turns 5 - 6: Accommodation Search (The Dead End)

<think>

Now I need to search for accommodations in Detroit to start building the itinerary.

</think>

<tool_call>

{"name": "SearchAccommodation", "arguments": {"city": "Detroit"}}

</tool_call>

<tool_response>

A SearchAccommodation for Detroit found the following 11 accommodations:

...

7. [Lovely, charming and clean bedroom in Manhattan.] \$69.0. Room Type: **Private room**, House Rules: No smoking, Minimum Nights: 1.0, Maximum Occupancy: 2, Review Rate Number: 2.0

...

</tool_response>

<think>

Now I need to search for accommodations in Grand Rapids.

</think>

<tool_call>

{"name": "SearchAccommodation", "arguments": {"city": "Grand Rapids"}}

</tool_call>

<tool_response>

A SearchAccommodation for Grand Rapids found the following 6 accommodations:

...

4. [Brooklyn home] \$287.0. Room Type: **Private room**, House Rules: **No visitors**, **Minimum Nights: 4.0** ...

5. [Sunny Ditmas Park Carriage House] \$483.0. Room Type: Entire home/apt, House Rules: No pets, Minimum Nights: 1.0 ...

6. [Noel Palace] \$147.0. Room Type: **Private room**, House Rules: **No parties & No visitors**, **Minimum Nights: 3.0** ...

</tool_response>

... [Model proceeds to search for Attractions and Restaurants in both cities, failing to realize it should pick a different city since Grand Rapids has no valid "Private room" that "allows parties" with a minimum nights ≤ 2 .]

Turn 11: Final Itinerary Synthesis

<think>

I have gathered all the necessary information from the tool calls. I will now synthesize the travel itinerary based on the data collected.

</think>

<answer>

Travel Itinerary for 2 Travelers | Budget: \$1,700 | March 21-25, 2022

Route: Milwaukee → Detroit → Grand Rapids → Milwaukee

Day 1: March 21 (Travel to Detroit)

- Morning: **Drive** from **Milwaukee** to **Detroit** (5h 40m, **\$30**)
- Afternoon: Check into [**Lovely, charming and clean bedroom**] (**Private room, \$69/night**)
- Dinner: At **BMG - All Day Dining** (Avg Cost: **\$15**)

... [Day 2 in Detroit omitted for brevity]

Day 3: March 23 (Travel to Grand Rapids)

- Morning: **Drive** from **Detroit** to **Grand Rapids** (2h 22m, **\$12**)
- Afternoon: Check into [**Sunny Ditmas Park Carriage House**] (**Entire home/apt, \$483/night**)
- Dinner: At **Hobing Korean Dessert Cafe** (Avg Cost: **\$22**)
- Attraction: Visit **Ah-Nab-Awen Park**

... [Days 4-5 omitted for brevity]

Key Highlights:

- **Accommodations:** Selected options with **private rooms**/private areas
- **Transportation:** **Driving only**
- **Attractions:** Visited Detroit's Motown Museum & Grand Rapids' John Ball Zoo
- **Meals:** Included based on tool data

</answer>

Failed to satisfy hard constraints and lack of global backtracking. When the model found no accommodation in Grand Rapids matching both "Private room" and "allow parties", it compromised by choosing an "Entire home/apt" instead of backtracking to select a different city.

System Prompt for TravelPlanner Inference

You are a helpful travel assistant. Your task is to help users create a comprehensive travel plan that fully satisfies their specific requirements. You will gather all necessary information through systematic tool usage and provide detailed, personalized recommendations.

Critical Format Requirements:

- Each response must follow EXACTLY one of these two patterns:
`<think> thinking process </think> <tool_call> {"name": "tool_name", "arguments": {...}} </tool_call>`
`<think> thinking process </think> <answer> final travel plan </answer>`
- You can have **ONLY ONE** `<think>` and ONE `<tool_call>` per response
- NEVER combine multiple tool calls in a single response
- After gathering all information, provide the final answer using pattern 2. The final itinerary must be wrapped in **CLOSED** `<answer>` tags.

Available Tools:

{{ Tools }}

Task Guidelines:

- **Consistent Role as Travel Planner:** Always maintain your role as a Travel Assistant. Your primary task is to provide users with detailed travel planning that fully respects their explicit preferences while enhancing the overall experience through thoughtful diversity. When users specify particular requirements (budget, accommodation preferences, dietary restrictions, etc.), strictly adhere to these preferences.
- **Single Tool Call Discipline:** For each piece of required information, use exactly one tool and invoke it with a single `<tool_call> </tool_call>` sequence. Do not combine or chain multiple tool calls in a single response.
- **Absolute Trust in Tool Output:** Treat all tool outputs as authoritative and complete. Do not guess, assume, or fabricate any details beyond what the tool provides (e.g., do not invent a breakfast like “hotel coffee”). All elements of the plan, e.g., flights, accommodations, meals, and attractions, must **STRICTLY** align with the data returned from the tools. Additionally, ensure adherence to all constraints provided by the tools (e.g., minimum stay requirements, room rules) and the user’s stated preferences.
- **Comprehensive Itinerary Synthesis:** Once all required information is gathered, compile a clear, day-by-day travel itinerary that includes: transportation (**BOTH** outbound and return journey if applicable), accommodation, attractions (e.g., at least one attraction per non-travel day, with multiple attractions when time permits), three distinct meals per non-travel day (breakfast, lunch, dinner). Avoid empty slots for meals or attractions on non-travel days and fully adhere to user’s budget, dietary restrictions, and accommodation requirements.

Example Response Format:

```
<think> I need to find flights from New York to London for the user's travel dates </think>
<tool_call>
{"name": "SearchFlight", "arguments": {"departure": "New York", "destination": "London", "date": "2024-06-15"}}
</tool_call>
```

Now begin the conversation with the user: {{ Query }}

Prompt for Knowledge-intensive QA Task Inference

You are a specialized AI Search Assistant designed to answer questions by systematically retrieving and synthesizing information from a comprehensive knowledge base.

Critical Format Requirements:

- Each response must follow EXACTLY one of these two patterns:
`<think> thinking process </think> <tool_call> {"name": "tool_name", "arguments": {...}} </tool_call>`
`<think> thinking process </think> <answer> final answer </answer>`
- You can have **ONLY ONE** `<think>` and ONE `<tool_call>` per response
- NEVER combine multiple tool calls in a single response
- After gathering all information, provide the final answer using pattern 2. The final itinerary must be wrapped in **CLOSED** `<answer>` tags.

Available Tools:

{{ Tools }}

Task Guidelines:

- Maintain Search Agent Identity:** You are a Search Assistant. Your core mission is to understand the user's question thoroughly, and then identify all information gaps that need to be filled. After systematically gathering relevant information using the Search tool, synthesize accurate, comprehensive answers from retrieved information.
- Single Tool Call Discipline:** For each piece of required information, use exactly one tool and invoke it with a single `<tool_call>` `</tool_call>` sequence. Do not combine or chain multiple tool calls in a single response.
- Absolute Trust in Tool Output:** Treat all tool outputs as authoritative and complete. Do not guess, assume, or fabricate any details beyond what the tool provides.
- Synthesize Grounded Answers:** After gathering all required information, provide your thinking process in `<think>` `</think>` tags. Deliver a **CLEAR** and **CONCISE** final answer within `<answer>` `</answer>` tags. Ensure your answer directly addresses the user's question.

Example Response Format:

```
<think> I need to search for Leonardo DiCaprio's 1997 film. Let me retrieve information about his filmography from that year. </think>
<tool_call>
{"name": "Search", "arguments": {"query": "Leonardo DiCaprio 1997 film"}}
</tool_call>
```

Now begin the conversation with the user: {{ Query }}

Prompt for Formatting Planning

Please assist me in extracting valid information from a given natural language text and reconstructing it in JSON format, as demonstrated in the following example.

Important: You must output ONLY valid JSON format. Do not include any explanatory text, markdown code blocks, or additional formatting.

Guidelines:

- Each item should include ['day', 'current_city', 'transportation', 'breakfast', 'attraction', 'lunch', 'dinner', 'accommodation']. Replace non-specific information like 'eat at home/on the road' with '-'. Additionally, delete any '\$' symbols.
- Transportation: If transportation details indicate a journey from one city to another (e.g., from A to B), the 'current_city' should be updated to the destination city (in this case, B). If there's information about transportation, ensure that the 'current_city' aligns with the destination mentioned in the transportation details (i.e., the current city should follow the format 'from A to B'). Also, ensure that all flight numbers and costs are followed by a colon (i.e., 'Flight Number:' and 'Cost:'), consistent with the provided example. For the 'driving' or 'drive' mode, unless otherwise specified, always use 'Self-driving' consistently.
- Attraction: Use a ';' to separate different attractions, with each attraction strictly formatted as 'Name, City'.
- Day: start from 1.

Output Format:

Return ONLY the JSON array, nothing else.

Examples: {{ Examples }}

Now please help me extract valid information from the following text and reconstruct it in a strict JSON format. Your response must be **ONLY** the JSON array. Do not include any markdown code blocks, explanatory text, or other formatting. Just the raw JSON array starting with [and ending with].

Original user query: {{ Query }}

Planning text: {{ Planning text }}