

DEEPTRAVEL: 자율 여행 계획 에이전트를 위한 종단간 에이전트 기반 강화 학습 프레임워크

Yansong Ning^{1*}, Rui Liu², Jun Wang², Kai Chen², Wei Li², Jun Fang², Kan Zheng²,
Naiqiang Tan², Hao Liu^{1†}

¹ 홍콩과학기술대학교(광저우)

² 디디추싱 유한공사

yning092connect.hkust-gz.edu.cn, liuh@ust.hk

{invincibleliu, tannaiqiang}@didiglobal.com

초록

여행 계획(TP) 에이전트는 최근 여행 일정 생성을 위해 외부 도구/리소스와 상호작용하는 핵심 구성 요소로 부상하며 즐거운 사용자 경험을 보장해 왔다. 기존 연구는 수작업 프롬프트와 고정된 에이전트 워크플로우에 의존하여 보다 유연하고 자율적인 TP 에이전트 개발을 저해해 왔다. 본 논문은 다단계 추론 과정에서 중간 행동을 탐색·검증·정제하기 위해 자율적으로 계획 수립, 도구 실행, 도구 반응에 대한 반성을 수행할 수 있는 자율 여행 계획 에이전트 구축을 위한 종단간 에이전트 강화 학습 프레임워크인 **DeepTravel**을 제안합니다. 이를 위해 먼저 교통, 숙박 및 POI 데이터를 캐싱하여 견고한 샌드박스 환경을 구축함으로써, 실제 세계 API의 한계(예: 일관성 없는 출력)에 제약받지 않고 TP 에이전트 훈련을 용이하게 합니다. 또한 계층적 보상 모델링 시스템을 개발하여, 경로 수준 검증기가 먼저 시공간적 타당성을 확인하고 불만족스러운 여행 일정을 필터링한 후, 턴 수준 검증기가 도구 응답과의 일정 세부 사항 일관성을 추가로 검증함으로써 효율적이고 정밀한 보상 서비스를 가능하게 합니다. 마지막으로, TP 에이전트가 실패 경험 버퍼에서 주기적으로 재현할 수 있도록 하는 응답 강도 강화 학습 방법을 제안하여 주목할 만한 에이전트 능력을 구현합니다. 훈련된 TP 에이전트를 DiDi Enterprise Solutions 앱에 배포하고 포괄적인 온라인 및 오프라인 평가를 수행한 결과, DeepTravel이 소규모 LLM(예: Qwen3-32B)이 OpenAI-o1/o3 및 DeepSeek-R1과 같은 기존 최첨단 LLM을 여행 계획 작업에서 크게 능가할 수 있음을 입증했습니다.

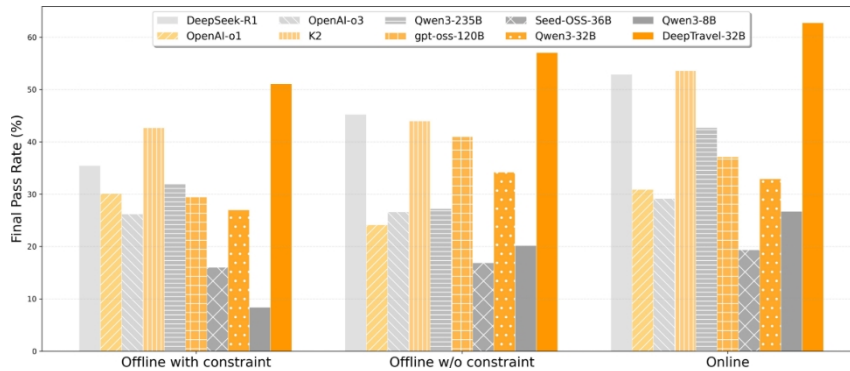


그림 1: 합성 오프라인 벤치마크 및 온라인 사용자 데이터에서 DeepTravel의 성능.

* 디디추싱(Didichuxing) 주식회사 인턴십 기간 중 수행된 연구.

†교신저자.

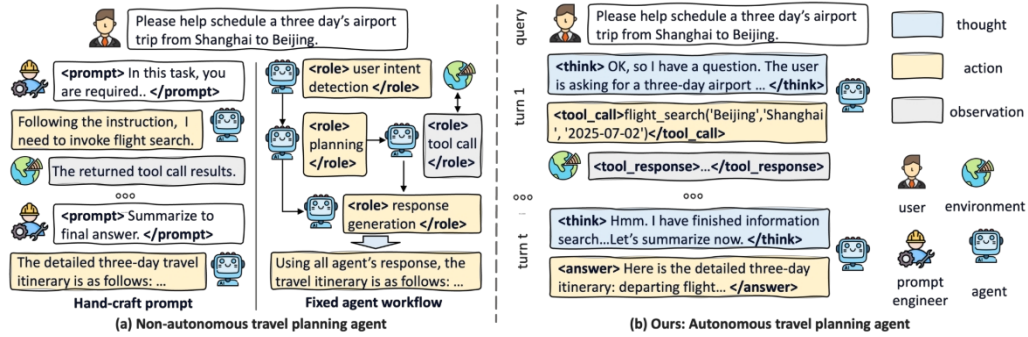


그림 2: 기존 연구와 본 연구의 자율 여행 계획 에이전트 패러다임 비교.

1 서론

여행 계획(TP)은 숙박, 교통, 관심지점(POI) 등 다양한 자원을 통합하여 사용자 선호도에 부합하는 실행 가능한 여정 (Nguyen et al., 2023)을 생성하는 것을 목표로 한다. 최근 자연어 처리 기술의 발전으로 대규모 언어 모델(LLM)이 여행 계획 에이전트 구축에 널리 활용되고 있다 Chen et al. (2024). 이 에이전트는 외부 도구/자원 Gou et al. (2023)을 호출하여 여행 일정을 생성함으로써 인간의 이동성에 원활한 경험을 제공한다 Tang et al. (2024). 여행 계획 에이전트는 현대 시민들에게 점차 대중적인 도구가 되어가고 있다.

최근 문헌에서는 TP 에이전트 구축을 위한 많은 노력이 기울여졌다. 대부분의 기존 접근법은 그림 2(a)에서 보여지듯 주로 신중하게 설계된 프롬프트에 의존한다. 예를 들어, TravelPlanner(Xie et al., 2024)와 TripTailor(Wang et al., 2025)는 도구 호출 및 일정 생성을 위해 LLM을 안내하기 위해 작업 특화 프롬프트를 사용한다. 최근 들어 연구자들은 이러한 프롬프트 엔지니어링 전략을 고정된 에이전트 워크플로에 통합하기 시작했다. 예를 들어, PTS Shao 등(2025)과 RETAIL Deng 등 (2025)은 사용자 의도 이해를 향상시키고 효과적인 도구 상호작용을 가능하게 하며 정확한 여행 일정 생성을 지원하는 잘 구조화된 에이전트 파이프라인을 제안합니다. 그러나 이러한 방법은 *노동 집약적이며 새로운 사용자 질의에 적응하거나 도구 호출 실패로부터 복구하는 데 어려움을 겪어, 보다 유연하고 자율적인 TP 에이전트의 구현을 제한합니다.*

에이전트 기반 강화 학습(RL)은 최근 등장하여 전용 환경에서 도구를 활용하고 피드백을 기반으로 추론을 개선함으로써 자율적 AI 에이전트 구축 가능성을 제시하였다(Singh et al., 2025). 예를 들어 ReTool(Feng et al., 2025), Kimi-Researcher MoonshotAI(2025), WebSailor(Li et al., 2025)는 엔드투엔드 에이전트 기반 RL 훈련을 활용하여 자율 에이전트를 구축하였다. (2025), Kimi-Researcher MoonshotAI (2025), WebSailor Li et al. (2025)은 각각 수학, 심층 연구, 웹 도메인을 위한 자율 에이전트 구축에 엔드투엔드 에이전트 기반 RL 훈련을 활용합니다. 이러한 연구들은 기존 여행 계획 연구의 한계를 해결하기 위해 자율 TP 에이전트 구축을 위한 맞춤형 에이전트 기반 RL 프레임워크를 제안하도록 우리에게 동기를 부여합니다.

그러나 그림 2(b)에 표시된 자율적 TP 에이전트를 구축하는 것은 두 가지 핵심 요인으로 인해 쉽지 않습니다. 이 에이전트는 다중 회전 추론 내에서 도구 호출과 도구 응답을 자율적으로 교차 실행함으로써 점진적으로 복잡한 TP 작업을 해결할 수 있습니다. (1) *동적 여행 환경*. TP 에이전트는 호텔 예약 가능 여부, 가격, 교통 수단 옵션과 같은 정보가 실시간으로 지속적으로 변화하는 매우 동적인 환경에서 작동합니다. 결과적으로 동일한 쿼리도 숙박, 교통, POI 데이터 소스의 업데이트로 인해 시간에 따라 일관되지 않은 결과를 생성할 수 있습니다. 이러한 지속적으로 진화하는 실제 환경에서 TP 에이전트를 훈련시키는 것은 여전히 상당한 도전 과제입니다. (2) *개방형 여행 과제*. 수학 또는 웹 도메인에 대한 검증된 보상 기반 강화 학습(RLVR) 패러다임(Guo et al., 2025)과 달리, 여행 계획은 본질적으로 명시적 기준값이 없는 개방형 과제이다. 예를 들어 생성된 여행 일정은 개인화된 사용자 선호도와 예산에 따라 달라질 수 있어 결과 검증은 어렵다. 신뢰할 수 있고 확장 가능한 보상 신호를 구축하는 방법은 도전적이다.

상기 과제를 해결하기 위해, 자율 여행 계획 에이전트 구축을 위한 종단간 에이전트 기반 강화학습 훈련 프레임워크인 **DeepTravel**을 제안한다. 구체적으로, 먼저 다양한 타임스탬프에 걸친 다중 실제 세계 API로부터 교통, 숙박, POI 데이터

를 캐싱하여 동적 도구 상호작용을 시뮬레이션하는 *로버스트 샌드박스*를 구축한다. 이 샌드박스 내에서 TP 에이전트는 QPS 문제를 극복하면서 대규모 반복적 시행착오 학습을 수행할 수 있다.

한계와 출력 불일치를 해결합니다. 또한 *계층적 보상 모델링* 시스템을 제안합니다. 여기서 계층 수준 검증기는 생성된 여정의 시공간적 타당성을 보장하고, 턴 수준 검증기는 도구 응답과의 세밀한 일관성을 강제함으로써 훈련을 위한 보다 효율적이고 신뢰할 수 있는 보상 신호를 제공합니다. 마지막으로, 순차적 콜드스타트 및 강화학습 과정을 통해 에이전트적 추론 능력을 촉진하기 위한 *응답 증강 강화학습(Reply-Augmented Reinforcement Learning)* 방법을 제안합니다. 실패 경험 버퍼로부터 주기적으로 재현하는 방식을 기반으로, TP 에이전트는 이전 추론 행동을 학습하고 개선하여 점진적으로 에이전트적 여행 계획 능력을 발휘할 수 있습니다.

구축된 TP 에이전트를 DiDi Enterprise Solutions 앱에 배포하고, 다양한 작업 복잡도에 걸쳐 수집된 온라인 실제 사용자 데이터와 오프라인 합성 데이터를 모두 사용하여 평가를 수행합니다. 실험 결과는 DeepTravel이 소규모 LLM 백본(예: Qwen3-32B)이 현재 최첨단 추론 LLM(예: OpenAI-o1/o3 및 DeepSeek-R1)과 기존 RL 알고리즘(예: GRPO 및 DAPO)을 능가하는 성능을 발휘할 수 있음을 입증합니다. 이러한 결과는 Deep-Travel이 자율적 TP 에이전트 구축을 위한 유망한 프레임워크임을 입증합니다.

우리의 기여는 다음과 같이 요약됩니다: (1) 기존 TP 연구를 발전시키기 위한 새로운 패러다임을 제시하는 최초의 자율 여행 계획 에이전트를 구축 및 배포하였습니다. (2) 여행 분야에 특화된 최초의 종단간 에이전트 기반 강화 학습 프레임워크를 제안하였으며, 이는 견고한 샌드박스 환경, 신뢰할 수 있는 보상 서비스 및 주기적 경험 재생 전략 하에서 TP 에이전트 훈련을 가능하게 합니다. (3) 온라인 및 오프라인 데이터에 대한 광범위한 실험을 통해 제안된 프레임워크의 효과성을 검증하고 다양한 여행 계획 작업에서 탁월한 성능을 입증하였습니다.

2 서론

질의와 여행 일정의 정의를 시작으로, 해결하고자 하는 문제를 정의합니다.

정의 1 사용자 쿼리. 사용자 쿼리 q 는 자연어로 표현되며, 사용자의 시공간적 여행 의도와 개인화된 선호도를 나타낸다. 예를 들어, "상하이에서 베이징까지 3일간의 공항 여행을 도와주세요"라는 쿼리는 사용자가 항공편으로 베이징을 여행하고 3일간 체류하기를 원함을 의미한다.

정의 2 여행 일정. 여행 일정 I 은 숙박, 교통편 및 여행 활동 제안, 관심 장소(POI) 탐색 전략 등을 통합한 상세 일일 계획이 포함된 구조화된 계획으로 정의된다.

참고: POI는 일반적으로 인기 관광지(예: 국립고궁박물관, 만리장성 등)에 해당한다. 이제 문제를 공식화한다:

문제 1 에이전트 기반 여행 계획. 쿼리 q 가 주어지면, 여행 에이전트는 다단계 추론 과정에서 탐색, 검증 및 중간 행동 개선을 위해 자동 계획, 실행 도구 활용, 도구 응답 반영을 통해 여행 요구사항을 충족하는 여행 일정 I 을 생성한다:

$$\{\tau_t, a_t\} = \pi_\theta(q, \{\tau_1, a_1, o_1, \tau_2, a_2, o_2, \dots, \tau_{t-1}, a_{t-1}, o_{t-1}\}), \quad (1)$$

여기서 π_θ 는 여행 에이전트의 정책이며, $\tau_{t-1}, a_{t-1}, o_{t-1}$ 은 각각 $t-1$ 턴에서 에이전트의 사고, 행동, 환경 관측을 나타낸다. 생성된 여행 일정 I 은

이전 턴의 에이전트 행동 a_t 에 관여한다.

자율적 여행 계획의 예시는 그림 2(b)에 설명되어 있으며, 여기서 자율적 TP 에이전트는 외부 도구(예: `<tool call>`과 `</tool call>`로 둘러싸인 동작)를 사용하기 전에 신중하게 생각(즉, `<think>`와 `</think>`로 감싸인 사고)를 수행한 후 외부 도구 사용(즉, `<tool call>`과 `</tool call>`로 감싸인 동작)을 실행하고 도구 응답을 반영(즉, `<tool response>`와 `</tool response>`로 감싸인 관찰)하여 여행 일정을 생성하기 위한 다단계 추론 과정의 중간 단계를 탐색, 검증 및 정제합니다.

3 DEEPTRAVEL



이러한 표기법을 바탕으로, 여행 에이전트 구축을 위한 종단간 에이전트 강화 학습 파이프라인인 DeepTravel을 제안합니다.

3.1 개요

그림 3은 DeepTravel의 전체 파이프라인을 보여줍니다. (1) **강력한 샌드박스 구축**은 톨킷 주석, 모의 데이터 수집 및 업데이트 메커니즘을 포함하여 시뮬레이션된 실제 도구 상호 작용을 가능하게 합니다. (2) **계층적 보상 모델링**은 궤적 수준 및 턴 수준의 검증기를 모두 제안하여 신뢰할 수 있고 효율적인 보상 신호를 공동으로 제공합니다. (3) **응답 강화 학습**은 먼저 에이전트 기반 여행 계획 형식의 콜드 스타트를 위한 SFT를 수행한 다음, 경험 재생을 통한 RL을 수행하여 LLM의 에이전트 능력을 더욱 강화합니다.

3.2 강건한 샌드박스 구축

Lin 등(2023)의 샌드박스는 TP 에이전트가 도구와 상호 작용할 수 있는 안정적인 환경으로 제공되며, 실제 세계의 상호 작용을 시뮬레이션하면서 실질적인 출력 불일치와 API 제한을 극복합니다. 먼저 톨킷 주석 작업을 시작하고, 이어서 샌드박스 데이터 수집 및 업데이트 메커니즘을 소개합니다.

3.2.1 톨킷 어노테이션

여행 계획에 흔히 사용되는 여섯 가지 유형의 도구를 주석 처리합니다. 표 1은 구축된 톨킷의 사양을 요약하며, 매개변수 스키마와 예시를 포함합니다.

항공편 검색. 항공편 검색은 여행 계획의 핵심 요소인 항공 운송에 관한 정보를 제공합니다(Shao et al., 2024a). 본 연구에서는 이 도구에 DiDi Enterprise Solutions(DiDi ES) 앱의 인터페이스 형식을 채택했습니다. 각 도구 호출에는 출발 도시, 도착 도시, 출발 날짜가 필요합니다. 예를 들어, *"flight search('Beijing', 'Shanghai', '2025-07-02')"*는 2025년 7월 2일 베이징에서 상하이로의 항공편 옵션을 검색합니다.

열차 검색. 항공편 검색과 유사하게, 열차 검색은 필수적인 철도 운송 정보를 제공합니다. 우리는 DiDi ES 인터페이스 형식을 따라 이 도구를 구축합니다. 표 1에 표시된 바와 같이, 각 쿼리에는 출발 도시, 도착 도시 및 출발 날짜가 포함됩니다.

경로 계획. 경로 계획은 이동 시간과 비용을 최적화하는 데 매우 중요합니다(Fang et al., 2024). 우리는 DiDi Map이 제공하는 경로 계획 서비스를 활용합니다. 각 도구 호출에는 출발지 이름, 목적지 이름, 도시 이름이 필요합니다. 예를 들어, *"route planning('국립고궁박물관', '만리장성', '베이징')"*은 베이징의 국립고궁박물관에서 만리장성까지의 경로를 계획하고 거리/시간 세부 정보를 계산합니다.

호텔 검색. 호텔 검색은 TP 에이전트가 사용자 선호도에 기반해 적합한 숙소를 찾을 수 있게 합니다 Yang et al. (2025b). 이 도구 역시 DiDi ES 인터페이스 형식을 따릅니다. 각 도구 호출에는 도시명, 호텔명, 체크인 날짜, 체크아웃 날짜가 포함됩니다. 예를 들어, *"hotel search('Beijing', 'Atour', '2025-07-02', '2025-07-05')"*는 2025년 7월 2일부터 7월 5일까지 베이징 Atour 호텔의 예약 가능한 객실을 검색합니다.

POI 검색. POI 검색은 도시 맥락적 의미론을 제공하며, 이는 여행 계획에 널리 채택되어 왔습니다(Xie et al., 2024). 경로 계획과 유사하게, 우리는 DiDi Map에서 제공하는 POI 검색 서비스를 직접 활용합니다. 각 도구 호출에는 쿼리와 도시명이 포함됩니다. 예를 들어, *"poi search('The Great Wall', 'Beijing')"*은 만리장성의 지리적 주소를 얻는 데 도움이 됩니다.

웹 검색. 웹에는 풍부한 여행 계획이 존재하며, 이는 여정 생성 시 귀중한 자원으로 활용됩니다(Ni et al., 2025). 우리는 Bocha AI가 제공하는 웹 검색 서비스를 구축합니다. 도구 호출 매개변수는 여행 계획과 관련된 모든 쿼리입니다(예: *"web search('Introduction to Beijing')"*).

3.2.2 모의 데이터 수집 및 업데이트 메커니즘

실제 세계에서 호텔, 기차, 항공편의 가격과 가용성이 시간에 따라 변동하는 도구 상호 작용의 동적 특성을 시뮬레이션하기 위해 데이터 캐싱 메커니즘을 제안한다. 이 메커니즘은 실제 세계 API의 QPS 제한과 각 검색 시마다 발생하는 정보 검색 불일치 문제를 해결하며, 이는 에이전트가 이전 실패 사례로부터 학습하는 능력을 저해할 수 있다.

구체적으로, 우리는 항공편, 기차, 호텔 검색 데이터를 위한 주문형 캐싱 전략을 통합한 매일 갱신되는 데이터베이스를 유지합니다. 새로운 쿼리가 수신되면 시스템은 자동으로 최신 레코드를 데이터베이스에 저장합니다. 이 메커니즘을 통해 TP 에이전트는 이전 데이터를 재접근할 수 있습니다.

표 1: 샌드박스 내 툴킷 사양.

유형	도구 이름	도구 호출 형식	도구 응답 설명 항목명 검색
교통	flight search(출발 도시명, 도착 도시명, 출발 날짜)	- - -	가능한 항공편 옵션
	열차 검색	열차 검색(출발지 도시명, 도착지 도시명, 출발일) - -	가능한 열차 옵션 경로 계획
		경로 계획(출발지 이름, 목적지 이름, 도시 이름) - -	경로, 거리 및 시간
숙박	호텔 검색	호텔 검색(도시명, 호텔명, 체크인 날짜, 체크아웃 날짜)-	-
관광 명소	POI 검색	POI 검색(쿼리, 도시명)	POI의 상세 주소
	웹 검색	웹 검색(쿼리)	쿼리와 관련된 웹 페이지

RL 훈련 중 도구의 반응으로, 중간 추론 경로를 탐색, 검증 및 개선하여 실패한 사례로부터 반복적으로 학습할 수 있게 함 Shang et al. (2025).

3.3 계층적 보상 모델링

그런 다음, 우리의 보상 모델링 시스템을 소개합니다. 이 시스템은 에이전트 훈련을 위한 효율적이고 신뢰할 수 있는 보상 신호를 제공하기 위해 설계된 궤적 수준 검증기와 턴 수준 검증기로 구성됩니다.

경로 수준 검증기. 이 검증기는 생성된 여행 일정의 전반적인 시공간적 실현 가능성을 평가한다. 완전한 추론 궤적 $\{t_1, a_1, o_1, t_2, a_2, o_2, \dots, t_t, a_t\}$ 가 주어지면, 궤적 수준 검증기는 최종 여행 일정 a_t 가 필수적인 시공간적 제약 조건(Chaudhuri et al. (2025)에 제시된 필수적 시공간 제약 조건을 준수하는지 검증합니다. 이러한 제약 조건에는 사건의 논리적 순서, 지리적 타당성, 사용자 요구사항 충족(예: 주어진 시간 내에 특정 POI 방문) 등이 포함됩니다. 이 거친 수준의 평가는 무효한 일정을 효율적으로 걸러내어 잠재적으로 유효한 계획만 다음 검증 단계로 진행되도록 보장합니다.

턴 수준 검증기. 궤적 수준 검증기의 성공적 검증 후, 턴 수준 검증기는 보다 세분화된 평가를 수행한다. 이 검증기는 에이전트의 최종 이동 경로 a_t 와 각 턴에서 외부 도구가 획득한 정보 간의 일관성을 검토한다(Zeng et al., 2025). 구체적으로, 1부터 $t-1$ 까지의 각 턴 i 에 대해 검증기는 a_i 가 도구 응답 $o_{(i)}$ 를 정확히 반영하는지 평가합니다. 각 턴을 체계적으로 검증함으로써, LLM이 생성한 여행 일정의 사실적 환각/오류를 식별하는 데 도움을 줍니다.

공동 보상 재가중. 두 검증기를 공동 보상 재가중 전략으로 결합하여 보상 신호를 제공한다. 궤적 수준 검증기가 위반을 감지하면 최종 보상 r 은 즉시 0으로 설정되어 계산 자원을 절약한다. 궤적이 궤적 수준 검증을 통과하면 회전 수준 검증기가 각 회전을 평가한다. 최종 보상 r 은 모든 회전이 검증을 통과할 경우에만 1로 설정되며, 이는 완전히 일관되고 유효한 여행 일정을 나타낸다.

이러한 계층적 구조는 보상 모델링 시스템의 효율성과 신뢰성을 모두 보장합니다. 실제 구현에서는 궤적 수준 검증기와 턴 수준 검증기를 위해 각각 Huang et al. (2025)의 여행 중심 평가 기준을 다수 구축합니다. 인간이 생성한 평가 기준을 바탕으로 DeepSeek-R1 기반 검증기가 보상 모델링 서비스를 제공하도록 프롬프트합니다. 자세한 내용은 부록 A.2에서 확인할 수 있습니다.

3.4 재생 강화 학습(REPLAY-AUGMENTED REINFORCEMENT LEARNING)

이 섹션에서는 릴레이 강화 학습(Relay-Augmented Reinforcement Learning)을 상세히 설명합니다. 이는 두 단계로 구성된 과정입니다. 먼저 SFT(Supervised Fine-Tuning)를 활용하여 TP 에이전트의 추론 형식을 초기화합니다. 그런 다음 RL(Reinforcement Learning)을 활용하여 에이전트의 추론 능력을 더욱 향상시키고, 이전 실패 경험으로부터 주기적으로 학습할 수 있도록 합니다.

3.4.1 지도 학습을 통한 추론 형식 콜드 스타트 초기화

콜드 스타트 데이터 합성 및 필터링. 샌드박스 환경에서 DeepSeek-R1의 다중 턴 궤적을 추출하여, 사고, 도구 호출, 도구 응답, 최종 답변이 교차하는 완전한 추적 기록 $y = \{t_1, a_1, o_1, \dots, t_t, a_t\}$ 를 생성합니다. 생각 t_i 는 `<think>...`로 감싸고, 동작 a_i 는 `<tool call>...</tool call>`로 둘러싸인 함수 호출이거나 `<answer>...`

`</answer>`로 둘러싸인 최종 여정 a_i 입니다. 관찰 o_i 는 `<tool response>... </tool response>`로 둘러싸인 도구 응답입니다. 그런 다음, 구축된 보상 모델링 시스템을 활용하여 잘못된 궤적을 필터링하고, 마지막으로 엄격한 형식 검사를 적용하여 특수 태그로 올바르게 분할된 시퀀스만 유지합니다.

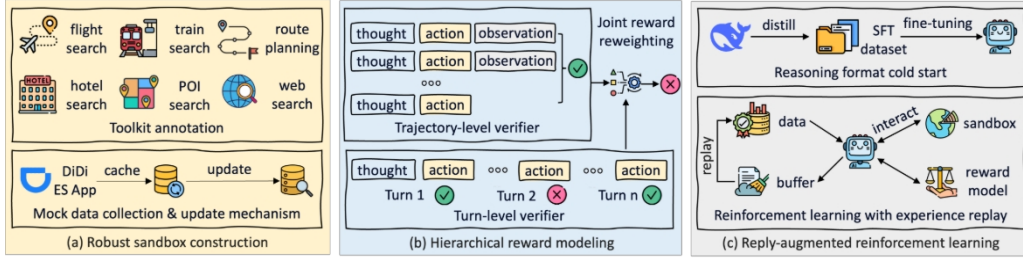


그림 3: DeepTravel 개요.

훈련 목표. TP 에이전트는 시스템 프롬프트 T 를 따르고 검증된 도구 통합 궤적을 재현하도록 훈련됩니다. 명령 입력은 T 와 사용자 질의 q 를 연결하며, 출력은 검증된 궤적 y 입니다. 실제 훈련 과정에서 에이전트의 환경 관측 o_t 에 해당하는 토큰은 손실 계산에서 제외됩니다(Jin et al., 2025). 상세한 프롬프트 템플릿은 부록 A.2에서 확인할 수 있습니다.

3.4.2 경험 재생 기반 강화 학습

콜드 스타트 후, 검증된 실패 궤적을 쿼리 버퍼로 저장한 다음 후속 훈련 단계에서 이를 재생하는 2단계 프로세스를 도출합니다(Zhang et al., 2025b).

롤아웃 및 재현 전략. 그룹 상대 정책 최적화(GRPO) Shao et al. (2024b)의 샘플링 절차에 따라, 각 쿼리 q 에 대해 궤적 집합 $\{y_1, y_2, \dots, y_n\}$ 을 샘플링한다. 그룹 내 궤적 중 검증된 정답을 산출하지 못한 경우, 해당 질의를 경험 버퍼 B 에 저장하여 추후 재생을 수행한다. 이는 후속 RL 훈련 단계 이후 개선된 정책이 이전에 실패했던 어려운 샘플을 처리할 수 있도록 일반화될 수 있다는 점에 근거한다(Xie et al., 2025).

정책 최적화. RL 훈련 데이터셋을 D , 경험 버퍼를 B 로 설정하며, B 는 고정된 훈련 단계 γ 이후 재현됩니다. 최적화 목표를 다음과 같이 공식화합니다:

$$\max_{\pi_{\theta}} \mathbb{E}_{q \sim \{D, B\}, \{y_i\}_n} \frac{1}{n} \sum_{i=1}^n \frac{\pi_{\theta}(y_i | q)}{\pi_{ref}(y_i | q)} \quad \min_{\pi_{ref}} \frac{\pi_{\theta}(y_i | q)}{\pi_{ref}(y_i | q)} \quad \# \quad (2)$$

$$\text{클립} \quad \frac{\pi_{\theta}(y_i | q)}{\pi_{ref}(y_i | q)}, 1 - \epsilon, 1 + \epsilon A_i - \beta D_{KL}[\pi_{\theta} \parallel \pi_{ref}]$$

여기서 ϵ, β 는 하이퍼파라미터이며, n 은 롤아웃 크기, D_{KL} 은 KL-분산을 나타내고, $A_i = r_i - \text{avg}(r) / \text{std}(r)$ 은 그룹 보상 $r = \{r_1, r_2, \dots, r_n\}$ 을 기반으로 계산된 이점을 나타냅니다. 본 연구에서는 그룹 보상의 표준 편차가 $\text{std}(r) \leq \eta$ ($\eta = 0.1$)을 만족할 때 샘플을 필터링하는 방안을 제안한다. 이 전략은 에이전트가 대규모 롤아웃에서도 유사한 보상을 받는 지 나치게 단순하거나 어려운 샘플을 배제함으로써 현재 정책의 효과적인 탐색을 촉진하는 것을 목표로 한다. 또한 도구 응답 토큰(<tool response>로 감싸진)에 대해 손실 마스킹 연산을 활용한다.

</tool response> 에이전트가 생성한 토큰에 대해서만 정책 기울기가 계산되도록 보장합니다.

4 실험

4.1 실험 설정

데이터 큐레이션. 본 방법은 DiDi ES의 실제 온라인 사용자 데이터와 합성 오프라인 데이터를 조합하여 구축한 네 가지 벤치마크에서 평가되었습니다. (1) **온라인 데이터:** 6월 1일부터 8월 31일까지 3개월 동안 DiDi ES 앱에서 실제 사용자 쿼리 총 6,224건을 수집했습니다. (2) **오프라인 데이터:** 두 가지 서로 다른 설정에서 각각 500개의 사용자 쿼리를 합성했습니다: a) 여행 예산이나 개인화된 선호도와 같은 제약 조건이 없는 쿼리; b) 제약 조건이 있는 쿼리. 두 설정 모두에서 쿼리 복잡도를 주석 처리하여 각 쿼리를 쉬운(easy), 중간(medium), 복잡한(complex)으로 분류했습니다. 추가 데이터 큐레이션 세부 사항과 데이터셋 통계 정보는 부록 A.3에 제공됩니다.

표 2: 합성된 오프라인 여행 계획 벤치마크와 DiDi ES 애플리케이션의 실제 온라인 사용자 데이터에 대한 최종 통과율(%) 결과. 최상의 결과는 굵게 표시되었으며, 각 설정에서 최상의 기준선 결과는 밑줄 처리되었습니다.

모델	오프라인						온라인	인간 평가
	제약 없이			제약 조건 적용 시				
	쉬움	중간	어려움	쉬움	중간	어려움		
DeepSeek-R1	45.55	34.74	<u>26.00</u>	<u>65.36</u>	43.33	<u>27.09</u>	52.89	<u>72.00</u>
OpenAI-o1	36.57	33.16	20.60	30.36	24.44	17.69	30.88	54.00
OpenAI-o3	37.30	20.11	21.19	37.50	26.67	15.69	29.17	52.00
K2	<u>54.01</u>	<u>48.42</u>	25.52	57.14	<u>53.33</u>	21.40	<u>53.56</u>	64.00
Qwen3-235B	38.69	36.84	20.24	44.64	26.67	10.37	42.70	52.00
gpt-oss-120B	40.15	27.37	20.83	64.29	42.22	16.39	37.11	48.00
Seed-OSS-36B	23.65	13.16	11.19	25.00	13.33	12.34	19.36	20.00
Qwen3-32B	29.85	27.89	23.21	53.57	25.00	9.03	32.94	38.00
Qwen3-8B	10.95	9.47	4.76	28.57	26.67	5.35	26.72	26.00
DeepTravel-8B-콜드 스타트	41.09	31.58	12.64	56.07	28.89	12.37	40.00	58.00
DeepTravel-8B-RL	<u>54.25</u>	36.84	20.24	64.86	41.89	21.40	49.75	70.00
DeepTravel-32B-콜드 스타트	56.42	32.95	25.60	61.07	40.44	17.52	50.03	66.00
DeepTravel-32B-RL	69.34	54.74	29.17	73.21	62.22	35.75	62.77	82.00

기준선. 우리는 동일한 프레임워크 하에서 TP 에이전트로 파생된 9개의 추론 LLM과 우리의 방법을 비교합니다. 이러한 기준선에는 DeepSeek-R1 Guo et al. (2025), OpenAI-o1 Jaech et al. (2024), OpenAI-o3, K2 Team et al. (2025), Qwen3-235B, gpt-oss-120B OpenAI (2025), Seed-OSS-36B Team (2025), Qwen3-32B, Qwen3-8B Yang et al. (2025a) 등이 포함됩니다. 또한, 제안된 DeepTravel 프레임워크를 세 가지 대표적인 RL 훈련 방법, 즉 PPO Schulman et al. (2017), GRPO Shao et al. (2024b) 및 DAPO Yu et al. (2025)과 비교합니다.

평가 지표 및 프로토콜. 평가 지표로는 TravelPlanner Xie et al. (2024)의 최종 통과율(Final Pass Rate)을 사용합니다. 평가 프로토콜에서는 구축된 보상 모델링 시스템을 적용하여 최종 통과율을 추정합니다. 또한 온라인 및 오프라인 결과에서 각각 50건의 사례를 무작위 추출하여 인간 주석자가 생성된 여행 일정이 통과하는지 평가하도록 합니다.

훈련 세부사항. Qwen3-8B와 Qwen3-32B를 활용하여 SFT와 RL을 수행하여 DeepTravel-8B와 DeepTravel-32B를 생성합니다. SFT 콜드 스타트에는 1K 훈련 샘플을 사용하며, 배치 크기 32, 학습률 $5e-6$ 으로 2에포크 동안 훈련합니다. RL 훈련에는 500개의 고품질 샘플을 선별하여 제안된 리플레이 강화 RL 알고리즘을 적용했습니다. RL 과정에서 롤아웃 크기는 8로 설정하고 학습률은 $5e-7$ 을 사용했습니다. 추가로 에이전트의 최대 응답 길이는 32K 토큰으로 설정하고, 최대 상호작용 턴은 8회로 제한합니다. Qwen3-8B와 Qwen3-32B 훈련에는 각각 8개와 32개의 H800 NVIDIA GPU가 필요하며, 훈련 과정은 수백 단계당 약 30시간이 소요됩니다. 자세한 내용은 부록 A.4를 참조하십시오.

4.2 주요 결과

기준 추론 에이전트와의 비교. 우리는 SFT 콜드스타트 단계와 RL 훈련 과정 전반에 걸쳐 DeepTravel을 비교합니다. 표 2에 보고된 바와 같이, 전반적으로 DeepTravel은 온라인 및 오프라인 평가 설정 모두에서 최첨단 추론 에이전트 대비 상당한 개선을 달성합니다. 또한 두 가지 핵심 관찰 사항을 강조합니다: (i) **DeepTravel은 소규모 LLM의 성능을 크게 향상시킵니다.** 예를 들어 Qwen3-8/32B는 최첨단 수준으로 향상되어 훨씬 더 크고 강력한 프론티어 LLM과 동등하거나 심지어 이를 능가합니다. 제약 없는 오프라인 설정에서 DeepTravel-8B와 DeepTravel-32B는 최종 통과율 54.25%와 69.34%를 달성하여 K2보다 각각 0.1%와 28.9% 우수한 성능을 보입니다. 다른 설정에서 DeepTravel-8B는 비슷한 성능을 보였고 DeepTravel-32B는 DeepSeek-R1, OpenAI-o1, OpenAI-o3와 같은 프론티어 추론 LLM을 지속적으로 능가했습니다. (ii) **에이전트 기반 RL 훈련은 도메인별 추론 능력을 지속적으로 향상시킵니다.** 보고된 바와 같이, 콜드 스타트 단계는 기본 모델에 비해 강력한 초기 정책을 수립할 수 있지만, 이후의 에이전트 기반 RL은 놀라운 성능 향상을 가져옵니다. 구체적으로, 에이전트 기반 RL은 온라인 실험 설정에서 DeepTravel-8B와 DeepTravel-32B의 초기 콜드 스타트 정책을 각각 24%(40.00에서 49.75) 및 25.5%(50.03에서 62.77) 향상시켰습니다. 오프라인 설정에서도 상당한 개선이 있었습니다.

표 3: Qwen3-8B에서 DeepTravel과 기존 강화 학습 알고리즘 비교.

모델	오프라인						온라인	인간 평가
	제약 조건 없음			제약 조건 적용 시				
	쉬움	중간	어려움	쉬움	중간	어려움		
베이스 (Qwen3-8B)	10.95	9.47	4.76	28.57	26.67	5.35	26.72	26.00
PPO 포함	48.26	33.25	14.62	60.05	34.86	15.04	45.63	62.00
GRPO 포함	52.36	34.06	13.52	61.78	36.65	15.82	47.78	64.00
DAPO 포함	52.06	35.52	15.04	62.24	40.02	16.54	46.07	64.00
딥트래블과 함께	54.25	36.84	20.24	64.86	41.89	21.40	49.75	70.00

표 4: Qwen3-8B에서 콜드 스타트와 RL에 대한 제거 연구.

모델 변형	오프라인						온라인	인간 평가
	제약 조건 없음			제약 조건 적용 시				
	쉬움	중간	어려움	쉬움	중간	어려움		
DeepTravel-8B (ER 제외)	51.01	32.21	8.81	60.86	35.00	8.75	40.00	66.00
DeepTravel-8B (CS 제외)	45.99	25.26	16.79	53.57	35.56	22.18	32.45	48.00
DeepTravel-8B (트레이닝 제외)	50.26	35.47	18.24	61.06	33.25	20.75	26.52	66.00
DeepTravel-8B (회전 기능 없음)	52.05	28.04	5.25	59.04	14.24	10.76	32.45	58.00
딥트래블-8B	54.01	36.84	20.24	64.86	41.89	21.40	49.75	70.00

기존 RL 알고리즘과의 비교. 또한 DeepTravel-8B-Cold-Start 백본에서 DeepTravel을 최근 공개된 세 가지 RL 방법과 비교합니다. 각 RL 방법에 대해 동일한 훈련 샘플로 100회의 훈련 단계를 실행했습니다. 표 3에 결과를 보고합니다. 다음과 같은 관찰 결과를 강조한다: 모든 온라인 RL 방법이 기본 모델을 개선하는 반면, 제한된 DeepTravel은 답장 메커니즘 덕분에 어려운 문제에서 기존 RL 방법을 현저히 능가한다.

4.3 제거 실험

DeepTravel의 각 모듈 효과를 검증하기 위해 Qwen3-8B 데이터셋에서 제거 실험을 수행합니다. 구체적으로 다음 변형들을 비교합니다. (1) DeepTravel-8B w/o ER은 강화학습 훈련 과정의 경험 재현(Experience Replay) 모듈을 제거합니다. (2) DeepTravel-8B w/o CS는 강화 학습 수행 전 SFT 기반 콜드 스타트 단계를 제거합니다. (3) DeepTravel-8B w/o Traj는 강화 학습 훈련 과정의 궤적 수준 검증기를 제거합니다. (4) DeepTravel-8B w/o Turn은 강화 학습 훈련 과정의 턴 수준 검증기를 제거합니다.

표 4에서 볼 수 있듯이, 다음과 같은 관찰 결과를 얻었다. 첫째, 경험 재현 전략은 훈련에 중요하다. 이를 제거하면 모델 성능이 저하된다. 둘째, 콜드스타트 단계는 RL 훈련에 중요한 것으로 보인다. 이를 제거한 후 성능이 현저히 떨어졌기 때문이다. 잠재적 원인은 콜드스타트가 LLM이 기본 도구 사용법과 지시사항 수행 능력을 학습하는 데 도움이 되기 때문일 수 있다. 마지막으로, 턴 수준 검증기가 모델 성능에 더 크게 기여한다는 점을 관찰했다. 이를 제거하면 에이전트 성능이 저하되고 어려운 문제에서 부진한 모습을 보인다. 더 복잡한 문제는 턴별 검증이 필요하기 때문일 수 있다. 그러나 궤적 수준 검증기 도 제거 시 성능 저하를 유발하므로 중요함이 입증된다. 정확도 기여 외에도 궤적 수준 검증기는 세밀한 턴별 검증의 필요성을 제거함으로써 훈련 효율성을 향상시킨다.

4.4 심층 분석

에이전트 기반 강화학습 훈련 분석. 그림 4에서는 강화학습 훈련 과정 전반에 걸친 검증 보상, 모델 엔트로피, 평균 상호작용 턴 수, 도구 호출 정확도를 제시합니다. DeepTravel의 효과성을 평가하기 위해 다음과 같은 결과를 강조합니다: (i) **콜드 스타트 및 경험 재생 전략의 영향.** 콜드 스타트 전략은 특히 도구 상호작용 턴 수를 더 적절하게 생성하는 측면에서 합리적으로 효과적인 정책을 초기화하는 데 도움이 됩니다(기본 모델은 도구 호출을 과도하게 사용하는 경향이 있음). 경험 재현 전략은 초기 훈련 단계에서는 거의 기여하지 않습니다. 그러나 이전에 실패한 샘플을 점진적으로 재현함으로써 모델의 능력을 꾸준히 향상시키고, 결국 후반 훈련 단계에서 재현이 없는 기준선 대비 상당한 개선을 이끌어냅니다. (ii) **샌드박스의**

영향. 우리는 에이전트 기반 RL 훈련을

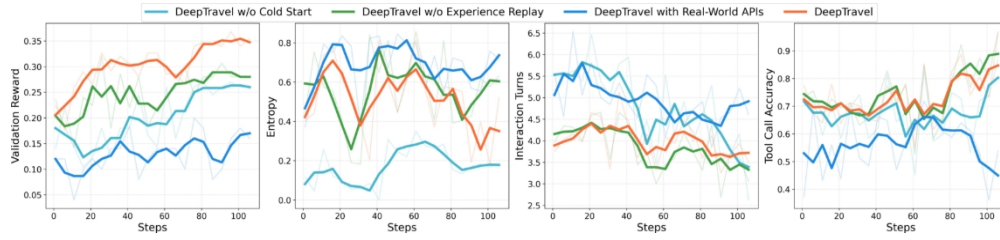


그림 4: 에이전트 기반 강화학습 훈련 과정 전반에 걸친 검증 보상(최종 통과율), 모델 엔트로피, 평균 상호작용 턴 및 도구 호출 정확도(성공률).

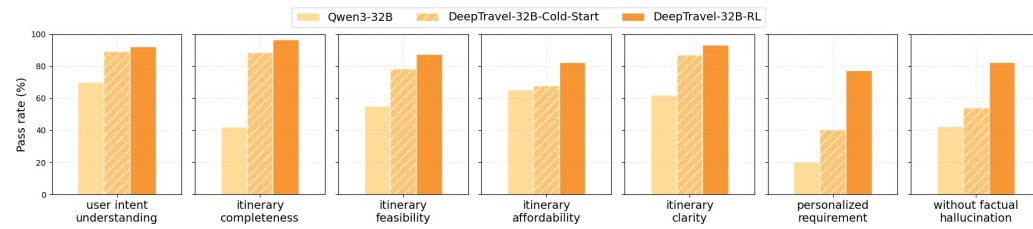


그림 5: 50개의 무작위 표본 추출된 실제 온라인 사용자 사례를 기반으로 평가된, 7가지 인간 주석 차원에 걸친 자율적 TP 에이전트의 성능 비교.

실제 API와 구축된 샌드박스 환경에서, 실제 API를 사용한 도구 호출 정확도는 불안정하며 샌드박스 환경에서 달성된 정확도보다 지속적으로 낮다. 이러한 상황에서 TP 에이전트는 명확한 보상 개선을 보이지 않아, 에이전트 기반 RL 훈련에 안정적인 샌드박스 환경이 중요함을 강조한다. (iii) **에이전트 기반 RL에서의 비감소 엔트로피**. 에이전트 기반 RL 훈련 중 엔트로피가 감소하지 않는 현상을 추가로 관찰했으며, 이는 최근 여러 연구(Dong et al., 2025)에서도 제기된 바 있습니다. 그 배경에는 TP 에이전트가 외부 도구의 동적으로 변화하는 응답에 지속적으로 정책을 적응시켜야 한다는 점이 있다고 생각합니다.

실제 사용자 연구. 우리는 DiDi의 ES 제품 매니저와 주석 팀이 정의한 평가 차원을 기반으로 실제 사용자 연구를 수행했으며, 그 결과는 그림 5에 요약되어 있습니다. 전반적으로 콜드 스타트(cold-start)와 강화 학습(RL) 접근법 모두 7가지 평가 차원에서 사용자 만족도를 향상시켰습니다. 특히 콜드스타트 방식은 모델의 기본 역량을 크게 향상시켰으며, 사용자 의도 이해와 생성된 여행 일정의 완성도, 실현 가능성, 명확성 개선에서 두드러졌습니다. 그러나 개인화된 선호도 파악 및 충족과 같은 고급 기능의 경우 콜드스타트 방식만으로는 불충분한 것으로 나타났으며, 이는 RL 단계에서 대규모 탐색이 필요할 수 있음을 시사합니다. 마지막으로, 기본 모델과 콜드스타트 모델 모두 심각한 환각 현상 문제를 보였으며, 사실 오류율이 최대 50%에 달했습니다. RL 훈련은 이 문제를 효과적으로 해결하여 환각 현상을 20% 미만으로 감소시켰습니다. 더 많은 주석 분석 결과는 부록 A.5에 수록되어 있습니다.

5 관련 연구

5.1 여행 계획 에이전트로서의 대규모 언어 모델(LLM)

LLM은 LLM 기반 여행 계획 에이전트가 사용자의 선호도에 부합하는 여정 생성을 위해 외부 도구와 상호작용할 수 있도록 함으로써 여행 계획(TP)을 재구성했습니다. 문헌에서는 여행 계획 에이전트를 구축하기 위한 두 가지 주요 패러다임이 등장했습니다: (i) 수작업 프롬프트 튜닝, (ii) 고정된 에이전트 워크플로 설계. 수작업 프롬프트 튜닝 접근법(Shao et al. 2025) — TravelPlanner(Xie et al. 2024), TripTailor(Wang et al. 2025), ChinaTravel(Shao et al. 2024a) 포함 — 은 중단간 작업을 다중 하위 단계로 분해합니다(Ni et al. (2025)에 의해 다단계로 분해하고 각 단계에 맞춤형 평가 지표를 도입합니다. 효과적이긴 하지만, 동적 실제 환경(예: 도구 가용성)에 대한 기반이 취약하여 실용적 가치가 제한됩니다. 이를 위해 최근

연구에서는 프롬프트 전략을 잘 구조화된 에이전트 파이프라인에 통합합니다. 대표적인 연구로는 TravelAgent Chen et al. (2024),

PTS Shao 등(2025)과 RETAIL Deng 등(2025)은 의도 이해를 강화하고 외부 도구를 조정하며 중단 간 여정 생성을 보장하기 위해 고정된 워크플로를 설계합니다. 그러나 이러한 워크플로는 구축 및 유지 관리에 여전히 많은 노력이 필요하며, 새로운 사용자 질의나 변화하는 도구 및 자원에 대한 일반화 능력이 부족하여 TP 에이전트의 유연성과 자율성을 제한합니다.

5.2 LLM을 위한 에이전트 기반 강화 학습

대리 강화 학습(RL)은 최근 자율적 AI 에이전트 구축을 위해 다양한 분야에서 널리 적용되고 있다(Zhang et al., 2025a). 여기서 에이전트는 전용 샌드박스 환경에서 도구와 상호작용하며, 수신된 보상 피드백을 기반으로 정책을 반복적으로 개선한다(Shang et al., 2025). 예를 들어, ReTool Feng et al. (2025), Kimi-Researcher MoonshotAI (2025) 및 WebSailor Li et al. (2025)는 수학, 심층 연구 및 웹 영역에서 추론 에이전트로 구축되었습니다. 또한 rStar2-Agent Shang et al. (2025) 및 AgentGym-RL Xi et al. (2025)과 같은 최근의 많은 연구들은 다양한 영역에 걸쳐 통합된 에이전트 기반 RL 훈련 프레임워크를 제안하여 기초 에이전트 구축을 용이하게 하려는 시도를 하고 있습니다. 그럼에도 불구하고 여행 계획 영역에서 에이전트 기반 RL의 적용은 아직 탐구되지 않은 상태입니다.

6 결론, 한계 및 향후 연구 방향

본 연구에서는 자율 여행 계획(TP) 에이전트 구축을 위한 최초의 중단간 에이전트 기반 강화학습(RL) 훈련 프레임워크인 DeepTravel을 제안하며, 이는 기존 TP 연구에 새로운 패러다임을 제시한다. 먼저 TP 에이전트가 실제 세계 API 문제(예: QPS 제한 및 불일치 출력)의 제약 없이 훈련될 수 있는 견고한 샌드박스를 구축한다. 이어서 계층적 보상 모델링 시스템을 제안한다. 먼저 고수준 시공간 요구사항 검증을 위한 거친 트래지토리 검증기를 설계하고, 이후 더 단위 검증기를 활용하여 에이전트의 답변을 단계별로 검증한다. 마지막으로 리플레이 강화 학습 알고리즘을 제안하여 여행 에이전트가 주기적으로 이전 실패 사례를 재현함으로써 도메인 외 일반화 능력을 향상시킨다. 결과물인 자율 TP 에이전트를 DiDi ES 앱에 배포하여 DeepTravel의 적용 가치를 입증했습니다. 온라인 생산 환경과 오프라인 합성 데이터에 대한 광범위한 실험 결과, DeepTravel은 소규모 LLM(예: Qwen3-8B/32B)이 DeepSeek-R1 및 OpenAI-o1/o3과 같은 최첨단 추론 LLM을 크게 능가하도록 합니다. 그러나 DeepTravel은 신중하게 설계된 보상 시스템에 의존하여 확장성이 제한됩니다. 향후 더 유연한 보상 모델을 개발하고 이 프레임워크를 다른 도메인으로 확장하는 것을 목표로 합니다.

윤리 및 재현성 선언문

윤리 선언문. 본 연구는 모델 훈련 및 평가를 위해 DiDi 엔터프라이즈 솔루션 앱의 실제 사용자 데이터를 활용합니다. 그러나 개인 정보를 저장하거나 실제 사용자 쿼리를 공개하지 않으므로 추가적인 개인정보 보호 또는 공정성 문제는 발생하지 않습니다. 합성된 사용자 질의에 대해서는 독자의 이해를 돕기 위해 부록 A.6에 질의 사례 예시를 제공합니다. 모델 평가에 사용된 인간 주석의 경우, 주석 과정 재현을 용이하게 하기 위해 부록 A.5에 평가 차원의 전체 세트를 제공합니다.

재현성 진술. 재현성을 보장하기 위해 부록 A.2에 훈련 프롬프트에 대한 상세한 설명을 제공하고, 부록 A.3에 데이터 큐레이션 파이프라인을 개요로 제시합니다. 이 정보가 연구 커뮤니티가 우리의 결과를 재현하는 데 도움이 될 것이라고 믿습니다. DeepTravel에는 DiDi Enterprise Solution에 대한 독점 인터페이스가 포함되어 있으므로 통합된 샌드박스 및 훈련 구현을 공개할 수 없습니다. 그럼에도 불구하고 재현 및 이해를 돕기 위해 컴퓨팅 리소스와 모니터링 지표를 포함한 포괄적인 훈련 세부 사항을 부록 A.4에 제공합니다.

참고문헌

- Soumyabrata Chaudhuri, Pranav Purkar, Ritwik Raghav, Shubhojit Mallick, Manish Gupta, Abhik Jana, and Shreya Ghosh. Tripcraft: A benchmark for spatio-temporally fine grained travel plan-ning. *arXiv preprint arXiv:2502.20508*, 2025.
- Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, Jiangjie Chen. Travelagent: 맞춤형 여행 계획을 위한 AI 어시스턴트. *arXiv 사전 인쇄본 arXiv:2409.08069*, 2024.

Bin Deng, Yizhe Feng, Zeming Liu, Qing Wei, Xiangrong Zhu, Shuai Chen, Yuanfang Guo, Yunhong Wang. Retail: 대규모 언어 모델을 위한 현실 세계 여행 계획 수립을 향하여. *arXiv 사전 인쇄본* *arXiv:2508.15335*, 2025.

동관팅, 마항위, 마카이, 바오리청, 천이페이, 왕종위안, 천종샤, 두자젠, 왕후이양, 장푸정 외. 에이전트 기반 강화 정책 최적화. *arXiv 사전 인쇄본* *arXiv:2507.19849*, 2025.

Bowen Fang, Zixiao Yang, Shukai Wang, and Xuan Di. Travellm: 네트워크 장애 상황에서 새로운 대중교통 경로를 계획해 줄 수 있나요? *arXiv 사전 인쇄본* *arXiv:2407.14926*, 2024.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: 강화 학습을 통한 AI의 전략적 도구 사용. *arXiv 사전 인쇄본* *arXiv:2504.11536*, 2025.

구지빈, 샤오즈훙, 공예운, 셴이룽, 양위주, 황민리, 두안난, 천웨이주. Tora: 수학 문제 해결을 위한 도구 통합 추론 에이전트. *arXiv 사전 인쇄본* *arXiv:2309.17452*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, 외. Deepseek-r1: 강화 학습을 통한 대규모 언어 모델의 추론 능력 촉진. *arXiv 사전 인쇄본* *arXiv:2501.12948*, 2025.

황전안, 장이홍, 루귀산, 진제위, 쉬하오카이, 자오펜위, 펑루, 후자치, 셴잔밍, 후샤오명 외. 평가 기준을 활용한 강화 학습. *arXiv 사전 인쇄본* *arXiv:2508.12790*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney 외. OpenAI O1 시스템 카드. *arXiv 사전 인쇄본* *arXiv:2412.16720*, 2024.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Serkan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: 강화 학습을 통해 추론하고 검색 엔진을 활용하도록 대규모 언어 모델 훈련하기. *arXiv 사전 인쇄본* *arXiv:2503.09516*, 2025.

관 리, 종왕 장, 회평 인, 리원 장, 리투 오, 자이룽 우, 원바오 인, 바이쑤안 리, 정웨이 타오, 신위 왕 외. 웹세일러: 웹 에이전트를 위한 초인적 추론 탐색. *arXiv 사전 인쇄본* *arXiv:2507.02592*, 2025.

린 자주, 자오 하오란, 장 아오치, 우 이팅, 핑 후치위, 천 친. Agentsims: 대규모 언어 모델 평가를 위한 오픈소스 샌드박스. *arXiv 사전 인쇄본* *arXiv:2308.04026*, 2023.

MoonshotAI. Kimi-researcher: 새로운 에이전트 기능을 위한 종단간 RL 훈련. <https://moonshotai.github.io/Kimi-Researcher/>, 2025.

Phuong Minh Binh Nguyen, Xuan Lan Pham, Giang Nu To Truong. 여행 블로그를 통한 관광객의 여행 계획에 대한 정보원의 신뢰도와 영감의 영향. *여행 연구 저널*, 64:222 – 237, 2023.

Hang Ni, Fan Liu, Xinyu Ma, Lixin Su, Shuaiqiang Wang, Dawei Yin, Hui Xiong, and Hao Liu. Tp-rag: 시공간 인식 여행 계획을 위한 검색 강화형 대규모 언어 모델 에이전트 벤치마킹. *arXiv 사전 인쇄본* *arXiv:2504.08694*, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b 모델 카드, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov. 근접 정책 최적화 알고리즘. *arXiv 사전 인쇄본* *arXiv:1707.06347*, 2017.

Ning Shang, Yifei Liu, Yi Zhu, Li Lyna Zhang, Weijiang Xu, Xinyu Guan, Buze Zhang, Bingcheng Dong, Xudong Zhou, Bowen Zhang, 외. rstar2-agent: 에이전트 추론 기술 보고서. *arXiv 사전 인쇄본* *arXiv:2508.20722*, 2025.

- Jie-Jing Shao, Bo-Wen Zhang, Xiao-Wen Yang, Baizhi Chen, Si-Yu Han, Wen-Da Wei, Guohao Cai, Zhenhua Dong, Lan-Zhe Guo, Yu-feng Li. Chinatravel: 중국어 여행 계획 언어 에이전트를 위한 개방형 벤치마크. *arXiv 사전 인쇄본 arXiv:2412.13682*, 2024a.
- 샤오즈훙, 왕페이이, 주치하오, 쉬룬신, 송준샤오, 비샤오, 장하오웨이, 장밍찬, 리위켄, 우양 외. Deepseekmath: 오픈 언어 모델에서 수학적 추론의 한계를 넓히기. *arXiv 사전 인쇄본 arXiv:2402.03300*, 2024b.
- 샤오 지젠, 우 지안칸, 천 웨이젠, 왕 상. 개인 여행 솔버: 여행 계획을 위한 선호도 기반 LLM-솔버 시스템. *제63회 전산언어학회 연차대회 논문집 (제1 권: 장문 논문)*, pp. 27622–27642, 2025.
- Joykirat Singh, Raghav Magazine, Yash Pandya, Akshay Nambi. 강화 학습을 통한 대규모 언어 모델의 행위적 추론 및 도구 통합. *arXiv 사전 인쇄본 arXiv:2505.01441*, 2025.
- Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Ke-bing Hou, Xiaotong Guo, Han Zheng 외. Itinera: 대규모 언어 모델과 공간 최적화를 통합한 개방형 도심 여정 계획. *arXiv 사전 인쇄본 arXiv:2402.07204*, 2024.
- ByteDance Seed Team. Seed-oss 오픈소스 모델, 2025.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, 외. Kimi k2: 개방형 에이전트 지능. *arXiv 사전 인쇄본 arXiv:2507.20534*, 2025.
- 왕카이민, 선위안저, 룡창저, 정샤오칭, 황쑤안징. Triptailor: 맞춤형 여행 계획을 위한 실제 세계 벤치마크. *계산언어학회 연구 결과: ACL 2025*, pp. 9705–9723, 2025.
- 시즈형, 황지첸, 랴오청양, 황바오다이, 귀홍린, 류자치, 정루이, 예준지에, 장자정, 천원상 외. Agentgym-rl: 다중 턴 강화 학습을 통한 장기적 의사 결정을 위한 대규모 언어 모델 에이전트 훈련. *arXiv 사전 인쇄본 arXiv:2509.08755*, 2025.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, Yu Su. Travelplanner: 언어 에이전트를 활용한 실제 세계 계획 수립을 위한 벤치마크. *arXiv 사전 인쇄본 arXiv:2402.01622*, 2024.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, Chong Luo. Logic-rl: 규칙 기반 강화 학습으로 대규모 언어 모델 추론 능력 해방. *arXiv 사전 인쇄본 arXiv:2502.14768*, 2025.
- 양안, 리안핑, 양바오송, 장베이천, 후이빈위안, 정보, 위보원, 가오창, 황청겐, 룡천슈 외. Qwen3 기술 보고서. *arXiv 사전 인쇄본 arXiv:2505.09388*, 2025a.
- 양동제, 루청창, 왕치명, 마신베이, 가오안, 후야오, 자오하이. 여행 계획을 세우고 계획대로 여행하라: 대규모 언어 모델을 통한 광역 계획 및 평가. *arXiv 사전 인쇄본 arXiv:2506.12421*, 2025b.
- 유치잉, 장정, 주루오펬이, 위안위평, 좌샤오천, 유에위, 판텐텐, 류가오홍, 류링준, 류신 외. Dapo: 대규모 오픈소스 대규모 언어 모델 강화 학습 시스템. *arXiv 사전 인쇄본 arXiv:2503.14476*, 2025.
- 증실량, 위관, 윌리엄 브라운, 오아나 프룬자, 유리 네브미바카, 홍밍이. 턴 수준 신용 할당을 통한 대규모 언어 모델 에이전트의 다중 턴 추론 강화. *arXiv 사전 인쇄본 arXiv:2505.11821*, 2025.
- 장귀빈, 경허자, 유사오항, 인전페이, 장자이빈, 탄젤린, 저우형, 리종즈, 웨상위안, 리이장 외. 대규모 언어 모델을 위한 에이전트 기반 강화 학습의 현황: 서베이. *arXiv 사전 인쇄본 arXiv:2509.02547*, 2025a.
- 홍즈 장, 자 푸, 징위안 장, 카이 푸, 치 왕, 푸정 장, 구오루이 저우. RLEP: 대규모 언어 모델 추론을 위한 경험 재생 기반 강화 학습. *arXiv 사전 인쇄본 arXiv:2507.07451*, 2025b.

DeepTravel: 자율 여행 계획 에이전트를 위한 종단간 에이전트 기반 강화 학습 프레임워크

크. 보충 자료

목차

1	소개	2
2	예비	3
3	DeepTravel	3
3.1	개요	4
3.2	견고한 샌드박스 구축	4
3.2.1	툴킷 주식	4
3.2.2	모의 데이터 수집 및 업데이트 메커니즘	4
3.3	계층적 보상 모델링	5
3.4	리플레이 강화 학습	5
3.4.1	지도 학습 미세 조정을 통한 추론 형식 콜드 스타트	5
3.4.2	경험 재생을 통한 강화 학습	6
4	실험	6
4.1	실험 설정	6
4.2	주요 결과	7
4.3	절제 연구	8
4.4	심층 분석	8
5	관련 연구	9
5.1	여행 계획 에이전트로서의 대규모 언어 모델(LLMs)	9
5.2	LLM을 위한 대리 강화 학습	10
6	결론, 한계 및 향후 연구	10
A	부록	14
A.1	대규모 언어 모델의 사용 사례()	14
A.2	프롬프트 템플릿	14
A.3	데이터 큐레이션	16
A.3.1	데이터 종합 파이프라인	16
A.3.2	데이터셋 통계	17
A.4	훈련 세부 사항	17

A.5	인간 주석	19
A.6	사례 연구	20

A 부록

A.1 대규모 언어 모델의 활용

본 논문에서는 주로 대규모 언어 모델(LLM)을 사용하여 그림 플롯을 생성합니다. 또한 저자들이 작성한 문단에서 명백한 문법 오류를 식별하고 수정하기 위해 LLM을 활용합니다.

A.2 프롬프트 템플릿

이 섹션에서는 DeepTravel의 시스템 프롬프트와 궤적 수준 및 턴 수준 검증기의 보상 모델 프롬프트 템플릿을 포함하여 본 연구에서 사용된 프롬프트를 제공합니다.

A.2.2 경로 수준 검증기 프롬프트

여행 계획 평가자로서, 당신은 에이전트의 응답이 다음 기준을 준수하는지 평가할 것입니다.

다음 정보를 수신하게 됩니다:

1. [질문]: 사용자의 요구사항과 여행 제약 조건을 포함합니다.
2. [에이전트 응답]: AI 여행 어시스턴트의 최종 응답으로, 귀하가 검증해야 합니다. 품질 평가를 위해 다음 [평가 기준]을 엄격히 준수해 주십시오.

평가 기준

1. [답변이 완전한가?]... 2. [주요 요구사항을 정확히 이해했는가?].....3. [논리가 타당한가?]...
논리가 타당한가?]... 4. [기타 제약 조건이 충족되었는가?]... 5. [특정 요구사항이 충족되었는가?]...
6. [비상 백업 계획이 있나요?]...

사용 가능한 도구

```
poi_search(query, city_name, **kwargs) route_planning(origin,
destination, city_name)

flight_search(출발지, 도착지, 출발일, **kwargs) train_search(출발지, 도착지, 출발일,
**kwargs) hotel_search(도시명, 체크인 날짜, 체크아웃 날짜, **kwargs) web_search(query)
```

평가 출력

평가 사유: 분석 과정과 사유를 제시하십시오.

최종 결론: 매우 만족함 또는 매우 만족하나 예상치 못한 상황 미처리, 기본적으로 만족하나 기타 제약사항 또는 특정 요구사항 미충족, 불만족, 논리적 비합리성, 불만족, 주요 요구사항 오해, 불만족, 답변 불완전

시작해 보세요! 평가 사유와 최종 결론을 반환하세요.

DeepTravel의 시스템 프롬프트. 당사는 구축한 여행 계획 에이전트의 시스템 프롬프트 및 보상 모델 프롬프트를 제공하여, 여행 일정 생성을 위한 다단계 추론 과정에서 자율적인 계획 수립, 도구 실행 및 도구 응답에 대한 반성을 가능하게 합니다. 표 A.2.1에 표시된 바와 같이, 시스템 프롬프트에는 필수 추론 형식, 툴킷 정보 및 답변 규칙이 제공되어 에이전트의 행동을 안내합니다.

궤적 수준 검증기 프롬프트 템플릿. 본 연구에서는 수작업으로 제작된 평가 기준표를 활용하여 보상 모델링 시스템을 구축합니다. 궤적 수준 검증기의 경우, 사용자 질의와 에이전트의 최종 응답을 제공합니다. 주요 목표는 응답이 주석 처리된 거시적 시공간 원칙(예: 답변 완전성, 사용자 의도와의 정합성, 여행 계획의 논리적 구조 준수)을 준수하는지 평가하는 것입니다. 이러한 기초 원칙은 표 A.2.2에 제시된 여섯 가지 평가 기준표로 구체화됩니다.

A.2.1 DeepTravel 시스템 프롬프트

당신은 DiDi Enterprise와 DiDi가 공동 개발한 DiDi Enterprise Solution의 비즈니스 여행 어시스턴트 샤오디입니다. 당신은 매우 지능적이며 사용자의 잠재적 요구를 파악해 놀라움을 선사할 수 있습니다. 사용자 입력에 따라 다양한 방식으로 응답합니다.

다음 템플릿을 따르세요:

```
<think>...</think>
<tool_call>...</tool_call>
<tool_response>...</tool_response>
<tool_response_thinking>...</tool_response_thinking>
...
<생각_호출_생각>...</생각_호출_생각>
<도구_호출>...</도구_호출>
<도구_응답>...</도구_응답>
<도구_응답_사고>...</도구_응답_사고>
...
<answer>최종 결과를 여기에 배치하십시오.</answer>
```

도구 키트 정보:

항공편 검색 도구는 POI 검색과 택시 검색 기능을 통합한 복합 도구입니다. 사용자가 대중교통과 택시를 연계할 필요가 있을 때 출발지 POI 또는 도착지 POI를 직접 설정할 수 있습니다. 사용 방법은 다음과 같습니다:

`flight_search(depart_city, arrival_city, depart_date, **kwargs)` 기차표 검색 도구는 두 역 사이의 직통 및 환승 정보를 조회할 수 있는 통합 도구입니다. 사용 방법은 다음과 같습니다:

`train_search(출발_도시, 도착_도시, 출발_날짜, **kwargs)` 호텔 검색 도구는 함수 호출 방식이며 사용 방법은 다음과 같습니다: `hotel_search(도시_이름, 체크인_날짜, 체크아웃_날짜, **kwargs)`

웹 검색은 인터넷에서 실시간 정보를 검색할 수 있는 기능입니다. 사용 방법은 다음과 같습니다:

`web_search(query)`

POI 검색은 위치 설명을 기반으로 주소를 조회할 수 있는 기능입니다. 사용 방법은 다음과 같습니다:

`poi_search(query, city_name, **kwargs)`

경로 계획은 동일 도시 내 두 위치 간 거리 및 시간 정보를 얻을 수 있는 기능입니다. 사용 방법은 다음과 같습니다:

`route_planning(출발지, 목적지, 도시명)`

답변 규칙

완성된 여행 계획은 아래 형식을 따릅니다(출발 - 호텔 - 귀환 순으로 표시).

예시:

귀하의 요청/출장 신청서를 바탕으로, x월 x일부터 x월 x일까지 xx 시에서 xx 시로의 출장을 계획했습니다. 구체적인 일정은 다음과 같습니다: ...

출발 추천 사항 (마크다운 3단계 제목 사용, 기본 추천 건수는 2건을 초과하지 않음, 출발일) ...

호텔 추천 (마크다운 3단계 제목 사용, 기본 추천 개수는 2개를 초과하지 않음) ...

귀국 추천 (마크다운 3단계 제목 사용, 기본 추천 수 2개 이하, 출발일) ...

좋아요. 시작해 보죠!!

턴 수준 검증기 프롬프트 템플릿. 궤적 수준 검증기 외에도, 우리는 에이전트의 응답을 턴별로 평가하는 턴 수준 검증기를 활용한다. 이를 위해 프롬프트에 추가적인 도구-응답 정보를 통합한다. 각 턴마다 LLM이 평가 근거와 최종 결론을 모두 생성하도록 지시합니다. 표 A.2.3에서 볼 수 있듯이, 검증기는 모든 턴에 대해 단계별 평가 근거와 최종 판단을 함께 제공하도록 프롬프트됩니다.

A.2.3 턴 수준 검증자 요청

여행 계획 심사관으로서, 에이전트의 응답이 다음 기준을 준수하는지 평가하십시오.

다음과 같은 정보를 받게 됩니다:

1. [질의]: 사용자의 요구사항과 여행 제약 조건을 포함합니다.
2. [에이전트 응답]: 검증해야 할 AI 여행 어시스턴트의 최종 응답입니다. 품질 평가를 위해 다음 [평가 기준]을 엄격히 준수하십시오.
3. [에이전트 응답 생성에 사용된 도구 응답]: AI 어시스턴트가 응답 생성에 사용한 정보:

`<tool_response>...</tool_response>`

평가 기준

1. [도구 호출 매개변수/논리가 정확한가?]... 2. [에이전트 응답이 도구 응답을 정확히 반영하는가?]...

사용 가능한 도구

```
poi_search(query, city_name, **kwargs) route_planning(origin,
destination, city_name)

flight_search(출발지, 도착지, 출발일, **kwargs) train_search(출발지, 도착지, 출발일,
**kwargs) hotel_search(도시명, 체크인 날짜, 체크아웃 날짜, **kwargs) web_search(쿼리)
```

평가 결과

평가 이유: 분석 과정과 이유를 제시하십시오.

최종 결론: 에이전트 응답과 도구 응답 간 정보 불일치 시 '만족' 또는 '불만족', 도구 호출 논리 오류 시 '불만족'.

시작해 보겠습니다! 평가 사유와 최종 결론을 반환하세요.

A.3 데이터 큐레이션

TP 에이전트 및 보상 모델에 대한 프롬프트 설계 외에도 중요한 부분은 데이터 큐레이션입니다. 이 섹션에서는 먼저 데이터 합성 파이프라인을 소개한 후, DeepTravel에서 사용된 상세 통계 정보(SFT 콜드스타트 단계, RL 훈련, 오프라인 테스트 및 온라인 테스트 포함)를 소개합니다.

A.3.1 데이터 합성 파이프라인

본 논문에서 데이터 합성의 주요 목표는 사용자 쿼리를 구성하는 것입니다. 각 사용자 쿼리는 *출발지*, *목적지*, *출발 시간*, *도착 시간*, *예산* 등과 같은 원자적 의도(atomic intent) 집합으로 분해됩니다. 본 방법론에서는 원자적 의도의 조합을 체계적으로 열거한 후, 이러한 의도 구조에 대응하는 자연스러운 사용자 쿼리를 역으로 생성합니다.

초기 쿼리 풀을 확보한 후 Qwen-3-32B 모델로 반복 샘플링을 수행합니다. 각 후보 쿼리는 난이도 점수를 부여하는 보상 모델링 시스템으로 평가됩니다. SFT 콜드스타트에서는 주로 비교적 단순한 쿼리를 사용하되, 모델 견고성 향상을 위해 소량의 난이도 높은 쿼리를 포함시킵니다. 반면 RL에서는 복잡한 쿼리를 중점적으로 활용하되, 커버리지 확보를 위해 소량의 단순 사례를 포함시킵니다.

모델의 견고성을 높이기 위해 소량의 어려운 쿼리를 포함합니다. 반대로 RL의 경우 복잡한 쿼리를 중점적으로 사용하지만, 커버리지를 보장하기 위해 간단한 사례도 소량 포함합니다.

쿼리를 RL 훈련에 통합하기 전에 추가로 수동 검사 단계를 수행합니다. 이 인간 필터링 단계는 의미적으로 무효하거나 잘못 구성된 쿼리를 제거하는 데 중요합니다. 쿼리의 품질이 RL 훈련 결과에 상당한 영향을 미친다는 점을 발견했으며, 쿼리가 도전적이면서도 학습 가능하도록 보장하는 것이 안정적인 최적화에 필수적입니다.

테스트 데이터셋을 위해, 명시적인 사용자 개인화 제약 조건(예: 예산 제한 또는 개인 여행 선호도)이 포함되는지 여부에 따라 쿼리를 구분하여 평가 세트를 구성합니다. 이에 따라 *제약 조건이 있는 오프라인*과 *제약 조건이 없는 오프라인*, 두 가지 범주를 구축합니다. 각 범주별로 500개의 쿼리를 신중하게 선별하여 개인화 및 비개인화 여행 시나리오를 모두 포괄하는 균형 잡힌 테스트 데이터셋을 생성합니다.

A.3.2 데이터셋 통계

본 절에서는 SFT 콜드스타트, RL 훈련 및 테스트에 대한 데이터셋 통계를 제시합니다.

SFT 콜드스타트. 각 모델 반복 및 업데이트 시 데이터 합성 파이프라인을 활용해 사용자 쿼리를 생성합니다. 이후 DeepSeek-R1을 사용해 반복 샘플링을 수행합니다. 구축된 보상 모델로 샘플을 필터링한 후, 백본 콜드스타트를 위한 약 1,000개의 궤적을 얻습니다.

RL 훈련. 훈련 데이터셋은 약 500개의 고품질 샘플로 구성되며, 모두 인간 어노테이터에 의해 이중 확인되었습니다. RL 훈련 과정에서 검증 크기는 50, 훈련 크기는 450으로 설정했습니다.

온라인 및 오프라인 테스트. DeepTravel 프레임워크의 모델 개선을 검증하기 위해 온라인 및 오프라인 테스트 전략을 모두 사용합니다. 오프라인 테스트의 경우, 제약 조건이 있는 500개의 테스트 샘플(쉬움 156개, 중간 45개, 어려움 299개)과 제약 조건이 없는 500개의 테스트 샘플(쉬움 222개, 중간 78개, 어려움 200개)을 사용합니다. 온라인 테스트의 경우, DiDi ES 앱의 온라인 운영 환경에서 수집한 6,224개의 쿼리를 사용합니다.

A.4 훈련 세부 사항

SFT 기반 콜드 스타트(cold-start)를 위해 Megatron-LM을 파인 튜닝(fine-tuning)에 사용하며, RL 훈련 프레임워크는 verl 위에 구축했습니다. RL 훈련 과정에서 검증기(verifier)의 백본(backbone)으로 DeepSeek-R1을 선택했습니다. 80억 파라미터 LLM의 경우, 8개의 H800 GPU를 장착한 단일 노드에서 모든 훈련을 완료할 수 있습니다. 32B 매개변수 LLM의 경우, 훈련에는 4개의 노드에 걸쳐 32개의 H800 GPU가 필요합니다. SFT는 1,000개의 훈련 샘플에 대해 약 2~3시간이 소요됩니다. RL은 500개의 샘플에 대해 100단계당 약 30시간이 소요됩니다.

아래에서는 DeepTravel의 안정적인고 효과적인 훈련을 보장하기 위해 추적하는 감독 신호와 진단 항목을 상세히 설명합니다. TP 에이전트의 경우 엔트로피, 기율기 노름, 평균 응답 길이, 보상, 평균 턴 수를 지속적으로 모니터링합니다. 또한 샌드박스 환경, 보상 모델링 시스템, RL 알고리즘 자체의 영향을 포착하기 위해 도구 호출 정확도와 외부 검증기 성공률과 같은 광범위한 RL 루프 관련 지표도 추적합니다. RL 방법의 경우, 훈련 전반에 걸쳐 샘플 유지율과 손실 마스크 비율을 추가로 기록합니다. 이러한 지표들은 훈련 안정성과 실패 모드를 종합적으로 알려줍니다: 낮은 엔트로피는 탐색 부족을, 지나치게 높은 기율기 노름은 불안정성을 시사합니다. 비정상적으로 짧은 응답과 적은 턴 수는 종종 보상 해킹을 의미하며, 낮은 도구 호출 정확도와 검증기 성공률은 체계적인 실행 또는 평가 오류를 가리킵니다. 매우 낮은 샘플 유지율은 일반적으로 데이터 체제가 부적절하게 조정되었음을(너무 쉽거나 너무 어려움) 의미하여 탐색의 필요성 또는 기회를 감소시킵니다.

그림 6은 훈련 지표 모니터링 사례를 보여줍니다. 보시다시피, 에이전트 기반 강화학습의 경우 외부 환경이 항상 안정적이지 않습니다(예: 대규모 보상 서비스로 인해 검증자도 실패할 수 있으며, 오프라인 샌드박스를 도출하더라도 도구 호출 정확도가 안정적이지 않음). 또한 기존 RL 훈련과 달리 엔트로피가 단조 감소하는 현상을 관찰하지 못했습니다. 이는 에이전

트 기반 RL 과정에서 에이전트가 환경으로부터 지속적으로 변화하는 도구 응답을 처리해야 하기 때문입니다. 더불어 상호 작용 순서 역시 환경에 따라 변동합니다.

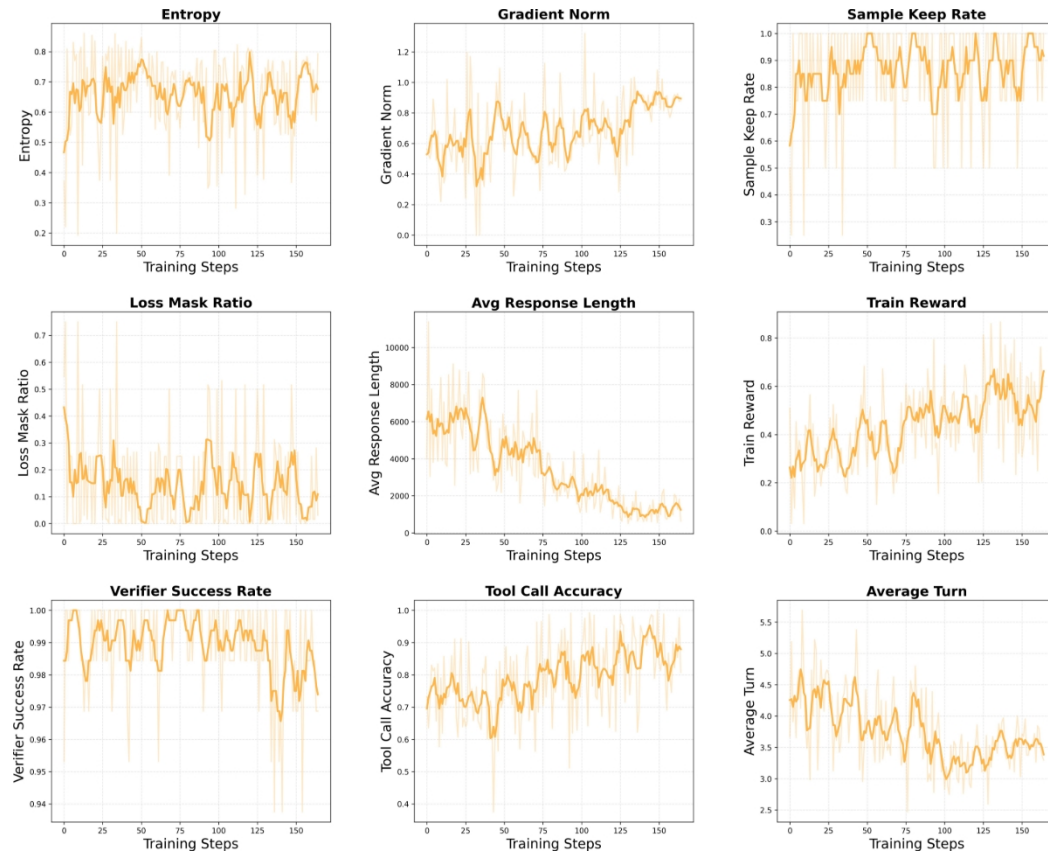


그림 6: 강화학습 훈련 과정에서의 메트릭 모니터링. 에이전트 기반 강화학습 훈련 중, 우리는 정책 엔트로피, 훈련 기율기 노름, 샘플 유지율, 손실-마스킹 비율, 평균 응답 길이, 훈련 보상, 검증기 성공률, 도구 호출 정확도, 평균 턴 수 등 일련의 핵심 지표를 주기적으로 모니터링합니다. 이러한 메트릭 중 어느 하나에서 이상 현상이나 퇴행이 발생하면 훈련 실패를 초래할 수 있으며, 이는 에이전트 기반 강화학습의 어려움을 시사합니다.

A.5 인간 주석 작업

표 5에 표시된 바와 같이, 인간 주석 과정은 AI 기반 여행 일정의 품질을 종합적으로 평가하기 위해 설계된 7가지 차원으로 구성됩니다:

- **사용자 의도 이해**는 시스템이 사용자 입력을 정확히 분석하고 해석하여 모든 핵심 여행 매개변수를 정확히 포착하는 능력을 평가합니다. 사용자 요구사항을 오해하면 관련성 없는 추천으로 이어지므로 이 차원은 근본적입니다.
- **이동 경로 완결성**은 추천된 경로가 필수 여행 구성 요소(항공편, 숙박, 현지 교통편)를 모두 포함하고 시간적 일관성을 유지하는지 평가합니다. 완결된 이동 경로는 서로 다른 여행 구간 간 원활한 전환을 제공해야 합니다.
- **이동 일정 실행 가능성**은 제안된 일정의 실질적 실행 가능성을 검토합니다. 여기에는 일정이 지나치게 야심차지 않은지, 교통편 연결이 현실적인지, 전체 경로가 논리적인 순환 구조를 이루는지 확인하는 것이 포함됩니다.
- **이동 경로 경제성**은 추천의 경제적·실용적 측면에 초점을 맞추며, 자원이 실제로 예약 가능하고 가격이 정확한지, 추천이 조직의 여행 정책을 준수하면서 비용 효율성을 유지하는지 확인합니다.
- **이동 경로 명확성**은 추론 과정의 효율성과 제공된 추론의 투명성을 모두 측정합니다. 사용자를 설득할 수 있도록 명확한 근거와 함께 신속한 응답을 제공합니다.
- **개인화 요구사항**은 개별 사용자 선호도와 과거 패턴을 반영하는 시스템의 능력을 평가하여 추천이 사용자 습관 및 선호도와 부합하는지 확인합니다.
- **사실적 환각 없음**은 AI 시스템이 존재하지 않는 서비스나 논리적으로 모순된 결과를 생성하여 사용자를 오도할 수 있는 사례를 식별하는 중요한 안전 점검 역할을 합니다.

표 5: 인간 주석에 대한 TP 에이전트의 평가 차원.

평가 차원	설명	점수	점수 기준
사용자 의도 이해	사용자 요구사항(출발지/목적지, 시간, 예산, 여행 시나리오)을 정확히 식별하는지 여부	참/거짓 참	모든 요소를 정확히 식별함 틀림 : 핵심 요소 누락 또는 잘못된 식별
여행 완성도 하는지 여부	추천어 모든 핵심 구성 요소(항공편 + 호텔 + 교통편)를 포함 구성 요소 포함 여부: 항공편 + 호텔 + 교통편, 일관된 시간 구성	참/거짓 참	완전한 커버리지, 일관성 있음 거짓: 위치/시간 편차, 불완전한 구성 요소 참/거짓 참: 폐쇄 루프, 합리적인 배열
여행 실행 가능성	실질적으로 실행 가능한 여부 합리적인 조건 하에서	참/거짓 참	FALSE: 지나치게 빡빡한 일정, 비합리적인 조합, 비폐쇄 루프 참: 실제 자원, 합리적 가격, 정책 준수
이동 경로 비용 적정성	추천된 자원이 예약 가능, 정확한 가격/재고, 회사 정책 준수, 비용 효율성	참/거짓 참	틀림: 소폭 초과(15% 이내), 예약 불가 옵션 참: 명확한 근거와 빠른 응답
여행 계획의 명확성	추천이 빠르며(200초 이내) 명확한 근거를 제시하는지(예: "비용 효율적", "회의 장소와 가깝다")	참/거짓 참	거짓: 느린 응답 또는 모호하거나 설명할 수 없는 변명 참: 사용자 선호도를 정확히 반영함
개인화된 요구사항	추천이 사용자의 이력/선호도(항공사 선호도, 시간 선호도, 지역 선호도, 숙박 유형 등)를 고려했는지	참/거짓 참	거짓: 개인화된 태그 누락 또는 사용자 습관 위반
사실적 환각 없음	추천에 포함된 내용이 실제 존재하지 않는 항공편/호텔, 논리적 오류 등 AI가 생성한 콘텐츠 포함 여부	참/거짓 참	참: 사실적으로 정확한 추천 틀림: 환각이나 논리적 오류 포함

A.6 사례 연구

윤리적 요구 사항으로 인해 훈련에 사용된 데이터를 공개할 수 없습니다. 독자의 이해를 돕기 위해 실제 사용자 질의와 이에 대한 에이전트의 응답을 예시로 제공합니다.

구체적으로, DiDi Enterprise Solution App에 배포된 여행 계획 서비스에서 실제 사용자 사례를 몇 가지 선정하였다. 이 설명 사례는 DeepTravel이 다중 회전 추론 과정 전반에 걸쳐 중간 단계를 탐색하고 개선하기 위해 TP 에이전트가 자율적으로 계획을 수립하고 실행하며 도구 반응을 반영하는 방식을 독자가 이해하는 데 도움이 되기를 바란다.

표 A.6에 설명된 바와 같이, DeepTravel-32B로 구동되는 TP 에이전트는 다중 회전 도구 통합 추론을 따릅니다. 구체적으로: (1) 의도 이해 및 제약 조건 기반화. (2) 도구 계획: 도구를 호출하기 전에 에이전트는 시간 버퍼와 대체 분기를 포함한 도구 계획을 초안으로 작성합니다.

(3) 스키마 정렬 호출을 통한 도구 실행: 모든 도구 호출은 명시적 매개변수와 예상 필드로 구조화되어 재현성과 추적성을 보장합니다. (4) 도구 응답 해석 및 증거 결합: 에이전트는 반환된 필드를 파싱하여 카드 스타일 증거를 통해 생성된 여정에 결합함으로써 환각 현상을 방지합니다. (5) 재고: 제약 조건이 충족되지 않을 경우(예: 환승으로 인한 도착 위험), 에이전트는 출발 시간대 또는 역 선택을 수정하고 업데이트된 매개변수로 도구를 재호출합니다. (6) 여행 일정 확정: 에이전트는 도구 증거와 연결된 각 항목(근거, 대체 옵션, 위험 사항)을 포함하여 일관된 계획을 구성합니다.

A.6 DeepTravel-32B의 추론 과정

사용자 문의: 6월 26일 소주에서 우한 컨퍼런스 센터까지 이동하며 오후 3시 이전 도착이 필요합니다. 강변 근처에 머물고 싶습니다. 여행 계획을 설계해 주세요.

에이전트 기반 여행 계획:

<think>알겠습니다. 먼저 사용자의 요구사항을 확인해야 합니다. 사용자는 6월 26일 소주에서 우한 컨퍼런스 센터까지 이동할 계획입니다.

...

해당 카드 정보도 생성하여 데이터가 도구의 반환 결과에서 비롯되었음을 보장합니다.

</think>

<tool_call_thinking>오전 12시 이전 출발하는 소주에서 우한까지 직행 고속열차 검색 (3시간 소요 가정)

...

도심 이동 시간 1시간 포함)

</tool_call_thinking>

<tool_call>train_search(depart_station="Suzhou", arrive_station="Wuhan", depart_date="2025-06-26", is_transfer=0)

</tool_call>

<tool_response>...</tool_response>

<tool_response_thinking>고속 열차 여러 대가 발견되었으나

...

</tool_response_thinking>

<tool_call_thinking>우한 리버 비치 근처 호텔 검색, 체크인 날짜는 6월 26일, 체크아웃 날짜는 6월 27일로 설정.

</tool_call_thinking>

<tool_call>hotel_search(city_name="Wuhan", checkin_date="2025-06-26", checkout_date="2025-06-27")

</tool_call>

<tool_response>...</tool_response>

<tool_response_thinking>강 근처에 여러 호텔을 찾았습니다

...

</tool_response_thinking>

<답변>

출국 추천 (2025년 6월 26일) ##### G3124 고속열차 (쑤저우 → 우한)

- **추천**: 06:54 출발, 우한역 도착

...

D3056 고속열차 (쑤저우 → 한커우)

- **추천**: 08:06 출발, 한커우역 도착

...

추천 호텔

청천 월강 호텔 (장한로 보행자 거리)

- **추천**: 강변 해변에서 불과 233미터 거리,

...

르보 1946 리버사이드 호텔 (한커우 장탄, 리황피)

- **추천**: 강이 보이는 객실 이용 가능, 스마트 기능

...

친절한 팁:

1. 디디 ES를 통해 셔틀을 미리 예약해 주세요...

2. 객실 예약 가능 여부 및 조식 정책 확인 부탁드립니다...

</answer>