**TripTailor系统技术文档**

**背景与研究动机**

随着大型语言模型（LLM）推理能力的提升，基于AI的旅行规划应用迅速兴起。旅行规划对个性化、高质量行程的需求不断增长[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=001%20The%20continuous%20evolution%20and,evaluation%20met%02010%20rics%2C%20which%20primarily)。然而，现有评测基准存在局限：例如**TravelPlanner**主要依赖模拟数据、**ChinaTravel**数据规模偏小，仅覆盖有限城市和POI[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=each%20user%20query,complexities%20of%20actual%20travel%20082)。这些基准无法真实反映LLM生成行程与人类设计行程的差异，现有评估指标也主要关注约束条件而忽略整体质量[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=001%20The%20continuous%20evolution%20and,evaluation%20met%02010%20rics%2C%20which%20primarily)。针对上述问题，TripTailor提出一个大规模、真实场景的旅行规划基准[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)。该基准收集了超过50万的真实兴趣点（POI）和近4000条详细旅行行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)，实验表明只有不足10%的LLM生成行程达到人类水平[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)，并揭示了可行性、合理性和个性化等多方面的挑战。

**系统架构**

TripTailor系统采用LLM驱动的多工具规划架构，对用户的旅行查询依次调用专门的工具并综合结果生成最终行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Figure%201%3A%20Overview%20of%20TripTailor,made%20travel%20plan)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=TripTailor%20provides%20a%20comprehensive%20sandbox,with%20user%20ratings%2C%20popularity%20201)。图1所示：给定用户查询，语言代理使用酒店查询、餐厅查询、景点查询、航班查询和火车查询等工具收集相关信息，逐步生成完整的行程计划[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Figure%201%3A%20Overview%20of%20TripTailor,made%20travel%20plan)。这些代理不仅要从多维度保证行程的可行性和合理性，还要深入分析用户特定需求和偏好，提供定制化方案[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Figure%201%3A%20Overview%20of%20TripTailor,made%20travel%20plan)。为此，我们构建了一个旅行规划沙盒环境，覆盖中国40个热门城市，内含航班和高铁时刻（约28,800条火车、15,100条航班）、5622个精选景点（含评分、门票、推荐时长等信息）、89,224家酒店和422,120家餐厅等数据[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=TripTailor%20provides%20a%20comprehensive%20sandbox,with%20user%20ratings%2C%20popularity%20201)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=indices%2C%20ticket%20prices%2C%20geographical%20locations%2C,For%20more%20206)。在规划过程中，代理根据查询调度各项工具，在该沙盒中检索对应的交通、住宿和景点信息，然后由LLM综合生成连贯的时间线行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Figure%201%3A%20Overview%20of%20TripTailor,made%20travel%20plan)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=TripTailor%20provides%20a%20comprehensive%20sandbox,with%20user%20ratings%2C%20popularity%20201)。

**数据集构建过程**

TripTailor数据集构建分为四个步骤，保障数据真实性和多样性[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=We%20collect%20information%20from%20the,In%20218)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=227%20ratings%20to%20ensure%20the,find%20suitable%20restaurants%20near%20the)：

* **Step I：沙盒环境搭建。** 从网络公开信息采集中国40城的景点、酒店和餐厅数据，以及城市间一周内的航班和火车时刻[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=We%20collect%20information%20from%20the,In%20218)。为了保证数据质量，只保留4A及以上级别或真实行程中出现过的景点，并补全其描述；剔除缺少价格、类别或评分的酒店和餐厅；对所有POI使用高德地图API补全经纬度信息[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=addition%2C%20we%20exclude%20hotels%20and,have%20received%20high%20226%203)。
* **Step II：真实行程构建。** 从在线旅行社获取自助游行程样本，选取内容详尽、评价较高的行程。随机指定每条行程的出发城市和日期，并使用沙盒中的航班/火车信息填充第一天和最后一天的交通细节。如果行程中未包含餐食，则为最近访问的景点附近自动推荐餐厅以补全餐饮安排[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=227%20ratings%20to%20ensure%20the,find%20suitable%20restaurants%20near%20the)。然后根据每日时间段从真实行程中提取景点列表，并从沙盒中补充其价格、类别和简介等属性。最后，调用LLM将这些原始信息重写为带时间轴的连贯详细行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=227%20ratings%20to%20ensure%20the,find%20suitable%20restaurants%20near%20the)。
* **Step III：用户查询生成。** 从每条行程中提取出发/返程日期、总时长、酒店星级以及餐饮和总预算范围等关键信息[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=244%20Step%20III%3A%20User%20Query,level)。将行程和这些要素输入LLM，要求其生成对应的用户查询语句。提示中使用第一人称口吻，并侧重描述高层次概念（如活动类型、景点主题、饮食偏好、行程强度等）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=244%20Step%20III%3A%20User%20Query,level)。生成的查询按难度划分：2-3天行程为“简单”，4-7天为“困难”[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=abstract%20concepts%20such%20as%20types,261)。
* **Step IV：质量控制。** 自动从行程文本中抽取关键信息，评估其可行性和合理性，对缺失或冲突的部分重新生成，如问题仍未解决则舍弃该样本[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Step%20IV%3A%20Quality%20Control,272)。同时，使用LLM检验生成查询与行程内容的一致性，并剔除不匹配的对。最终对测试集样本的关键信息进行人工校对，确保数据质量[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Step%20IV%3A%20Quality%20Control,272)。

以上步骤最终产生了包含查询–行程对的评测数据集。总体数据分布见附录：共40个城市、$\sim$50万POI，近4000个详细行程样本（包含交通、住宿、餐饮、景点等信息）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=227%20ratings%20to%20ensure%20the,find%20suitable%20restaurants%20near%20the)。

**三维度评估指标**

TripTailor从**可行性**、**合理性**和**个性化**三个维度对生成行程进行评估[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。具体指标包括：

* **可行性（Feasibility）**：检查行程中使用的信息是否全部来自预定义的沙盒环境，否则视为幻觉；同时要求行程包含所有必要元素（例如安排住宿）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)。若不满足则该行程标记为不可行。
* **合理性（Rationality）**：评估行程的逻辑与效率。指标包括餐厅多样性（全程不重复同一家）、餐饮价格是否在预算范围内；景点多样性（不同日程内不重复同一景点）；各景点的游览时长是否在推荐范围内；总费用是否超出预算；以及路线优化程度（尽量减少同日景点间的平均行驶距离，提高出行效率）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Diverse%20Restaurants%20Restaurant%20choices%20should,minimize%20travel%20time%20between%20POIs)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=recommended%20time%20range,habits%2C%20and%20other%20relevant%20factors)。
* **个性化（Personalization）**：衡量行程与用户兴趣爱好之间的契合度。重点考察行程是否融入了用户指定的主题元素，如景点类型、餐饮偏好、活动类型、行程节奏、消费习惯等[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。

在个性化评估上，我们采用了双重策略：一方面使用LLM作为“评判者”，同时将参考真实行程与生成行程输入LLM，让其对比酒店、景点、餐厅选择及旅行深度、广度等因素并评分[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=312%20%E2%80%A2%20LLM,it%20meets%20user%20preferences%20and)；另一方面训练奖励模型，通过对相似查询构造正负行程对进行学习，预测行程满足用户需求的程度[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=%E2%80%A2%20Reward,an%20exact%20match%2C%20while%20347)。最终量化指标包括可行率、合理率、个性化超越率、平均路线距离比和最终超越率等，用于直观比较LLM生成行程与真实行程之间的差距[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=,18)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=380%20%E2%80%A2%20Average%20Route%20Distance,meet%20personal%02391%20ization%20standards%20if)。

**实验设置与结果**

我们在TripTailor基准上评估了多种主流LLM及规划策略[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=LLMs,also%20develop%20a%20manual%20404)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=workflow%20decomposition%20method%20to%20serve,near%20these%20attractions%20411%206)：

* **模型**：测试了闭源模型（OpenAI GPT-4o、GPT-4o-mini）以及开源模型（DeepSeek-V3、o1-mini）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=LLMs,2022%29%2C%20and%20Reflexion%20403)。
* **规划方法**：包括直接生成（Direct）、零样本链式思维（CoT）、ReAct、反思（Reflexion）等自动策略[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=ning%20methods%3A%20Direct%2C%20Zero,between%20cities%2C%20after%20which%20407)；以及我们提出的\*\*工作流分解（Workflow）\*\*基线：逐步拆解行程规划问题——先确定城市间交通方案，再用LLM按用户偏好排序景点并选取重点，生成初步行程，最后结合景点地理位置完善餐饮安排[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=workflow%20decomposition%20method%20to%20serve,near%20these%20attractions%20411%206)。

主要实验结果如表2所示[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)：在TripTailor上，工作流分解策略在两日游任务中可行率达98.9%、合理率98.6%，但其生成行程的个性化超越率仅为74.3%（仍远低于人类水平）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。相比之下，直接生成方法（Direct+GPT-4o-mini）的可行率和合理率明显下降（分别为93.1%和87.0%），个性化超越率仅约11.6%[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。总体来看，所有方法生成行程的平均路线距离均显著高于真实行程：即使最优方法平均距离也达真实的1.8倍，一般方法更高[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。而“最终超越率”指标显示，通过满足所有可行性和合理性要求的生成行程中，只有极少部分在个性化方面超越了真实行程（大多数方法最终超越率不足20%）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=380%20%E2%80%A2%20Average%20Route%20Distance,meet%20personal%02391%20ization%20standards%20if)。这些结果表明当前LLM在满足约束、优化路径和捕捉个性化需求方面仍存在明显不足，远未达到人类规划的水平[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=5%20Conclusion%20540%20We%20introduce,548)。

**技术实现要点**

TripTailor系统在实现上融合了多项关键技术：

* **模型与工具**：使用多种LLM作为核心，配合专门的查询工具链。工具包括航班查询、高铁查询、酒店查询、餐厅查询和景点查询等，这些工具依托沙盒数据库进行信息检索[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=TripTailor%20provides%20a%20comprehensive%20sandbox,with%20user%20ratings%2C%20popularity%20201)。例如，通过航班/高铁工具获取交通方案，通过酒店/景点工具获取候选地点，并按条件过滤出最匹配项。
* **提示设计**：针对不同策略设计了定制提示。例如在CoT、ReAct等策略中，引导模型分步骤推理行程安排；在用户查询生成时，提示模型使用第一人称陈述需求，并突出活动类型、景点类别、预算等高层信息[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=244%20Step%20III%3A%20User%20Query,level)。在工作流分解策略中，则先提示模型确定交通方案，再提示其对景点进行偏好排序，最后生成完整行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=workflow%20decomposition%20method%20to%20serve,near%20these%20attractions%20411%206)。
* **流程分解策略**：我们的工作流分解方法仿照人类规划思路，对规划过程进行手动拆分[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=workflow%20decomposition%20method%20to%20serve,near%20these%20attractions%20411%206)。该方法首先确定城市间可行的交通方式，然后使用LLM为候选景点排序并选取重点景点，生成初步行程，再根据景点位置和餐厅信息完善用餐安排[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=workflow%20decomposition%20method%20to%20serve,near%20these%20attractions%20411%206)。
* **沙盒环境维护**：沙盒数据库使用外部数据源补充。例如，利用高德地图API获取景点和酒店的经纬度坐标[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=addition%2C%20we%20exclude%20hotels%20and,have%20received%20high%20226%203)。此外，为训练个性化奖励模型，我们使用TF-IDF检索技术，从用户查询库中找到与当前查询相似但不同的条目，构造正负样本对进行模型训练[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=%E2%80%A2%20Reward,an%20exact%20match%2C%20while%20347)。
* **验证与评估**：行程生成后，系统自动检查冲突或缺失（如预算超标、时间冲突等），必要时调用LLM或规则进行修正。同时，集成基于LLM和奖励模型的双重评估机制，对比生成行程与真实行程的差异，从多个维度验证行程质量[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=312%20%E2%80%A2%20LLM,it%20meets%20user%20preferences%20and)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=380%20%E2%80%A2%20Average%20Route%20Distance,meet%20personal%02391%20ization%20standards%20if)。

**与现有基准对比分析**

相较于先前基准，TripTailor在真实性、规模和评估维度上都有显著提升[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=each%20user%20query,complexities%20of%20actual%20travel%20082)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)。TravelPlanner主要使用模拟数据，难以反映实际场景[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=each%20user%20query,complexities%20of%20actual%20travel%20082)；ChinaTravel虽基于真实数据，但仅覆盖10个城市、每城约1200个POI，规模远小于实际需求[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=each%20user%20query,complexities%20of%20actual%20travel%20082)。TripTailor基于真实世界数据构建，包含$\sim$50万POI和$\sim$4000个详尽行程案例，覆盖40个重点城市[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)。同时，TripTailor评估不仅关注基本约束满足，还专门设计了度量行程合理性和个性化的新指标[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。通过更真实的大规模数据和更全面的评价体系，TripTailor为研究者提供了更具挑战性和代表性的旅行规划测试平台，推动自动规划代理的进一步发展[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=5%20Conclusion%20540%20We%20introduce,548)。

**TripTailor技术报告**

**研究目标与意义**

TripTailor旨在为个性化旅行规划任务构建真实可靠的评测基准，评估智能行程规划系统的性能和局限性。相比以往基于模拟或小规模数据的基准，TripTailor采用大规模真实数据集，为研究者提供更具挑战性的场景。通过引入新的可行性、合理性和个性化指标，TripTailor能够全面衡量AI规划器在满足用户需求和设计高质量行程方面的能力[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)。该基准将推动更智能的旅行规划代理研发，使其能够在实际应用中更好地理解并满足用户偏好。

**数据集规模与内容**

TripTailor数据集规模庞大、内容丰富[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)：

* **兴趣点（POI）数据**：$\sim$50万条，涵盖40个城市的景点、酒店、餐厅等，每个POI带有评分、热度、价格、简介和地理位置等详细属性[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=TripTailor%20provides%20a%20comprehensive%20sandbox,with%20user%20ratings%2C%20popularity%20201)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=indices%2C%20ticket%20prices%2C%20geographical%20locations%2C,For%20more%20206)。
* **真实行程样本**：近4000条，均为自助游用户行程，包含具体的出发/返回时间、交通方案、住宿安排、每日游览活动和餐饮信息[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=227%20ratings%20to%20ensure%20the,find%20suitable%20restaurants%20near%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)。
* **用户查询**：对应每条真实行程自动生成一个自然语言查询，包含出发地、出发/结束日期、预算、偏好景点类别、饮食偏好、行程强度等要素[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=244%20Step%20III%3A%20User%20Query,level)。

研究者使用时，可输入用户查询，系统将在沙盒环境中搜索匹配选项并生成行程。TripTailor提供行程生成和评估的完整流程及工具，便于对比不同模型在旅行规划任务中的表现[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)。

**系统使用流程示例**

1. **用户输入需求：** 例如，“成都出发到上海的2日游，周四上午出发，周五晚上返回，预算3100元；希望入住4星级左右的酒店，每餐预算50-100元；偏好海洋馆和野生动物园等主题景点”。
2. **交通方案检索：** 系统使用航班/高铁查询工具，在指定时间范围内选择合适的交通方式。例如找到成都07:45飞上海浦东、10:35到达的航班[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Query%2007%3A45%E2%80%9310%3A35%20,at%20Shanghai%20Wenfei%20Boutique%20Hotel)。
3. **住宿与景点检索：** 系统根据条件检索酒店（中档、4星级）和景点，如上海海昌海洋公园、上海野生动物园等符合海洋动物和野生动物主题的景点[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=13%3A00%E2%80%9317%3A00%20,last%20admission%20at%2016%3A00)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=09%3A00%E2%80%9312%3A00%20,exhibits%2C%20and%20animal%20behavior%20demonstrations)。
4. **行程生成：** LLM综合以上信息，按时间顺序生成详细行程。示例输出（简化版）：
   * **第1天**：07:45-10:35 乘坐CA4503航班成都至上海（票价¥870）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Query%2007%3A45%E2%80%9310%3A35%20,at%20Shanghai%20Wenfei%20Boutique%20Hotel)；11:30-12:00 入住上海文飞精品酒店（4星，¥327/晚）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=11%3A30%E2%80%9312%3A00%20%7C%20Check,level%20tourist%20attraction%20featuring%20five)；13:00-17:00 游览上海海昌海洋公园（4A级，门票¥79，海洋主题）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=11%3A30%E2%80%9312%3A00%20%7C%20Check,thrilling%20attractions%20such%20as%20Volcano)；18:00-19:30 在“在水一方·鱼上世界”火锅店用餐（¥81/人）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=18%3A00%E2%80%9319%3A30%20,0%2F5)。
   * **第2天**：09:00-12:00 参观上海野生动物园（4A级，门票¥165，动物园主题）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=09%3A00%E2%80%9312%3A00%20,exhibits%2C%20and%20animal%20behavior%20demonstrations)；12:00-13:00 园内简餐午餐（¥79.5）；13:00-14:00 前往机场；16:25-20:05 搭乘CZ3359航班返回成都（票价¥1130）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=animals%2C%20including%20giant%20pandas%2C%20giraffes%2C,Recommended%20duration%3A%203%20hours)。  
     （以上行程示例源自TripTailor系统输出[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Query%2007%3A45%E2%80%9310%3A35%20,level%20tourist%20attraction%20featuring%20five)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=18%3A00%E2%80%9319%3A30%20,exhibits%2C%20and%20animal%20behavior%20demonstrations)，仅供说明。）
5. **行程评估：** TripTailor使用预定义指标自动评估生成结果。可行性检查如验证交通和住宿安排是否合理、是否全部来源于沙盒数据；合理性检查如路线是否优化、费用是否超出预算；个性化检查行程是否反映用户兴趣偏好。研究者还可使用TripTailor提供的评估工具，将生成行程与真实行程进行对比，分析差异及不足[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。

**样例输出与评估方式**

TripTailor输出标准化的多日行程，每个活动项目包含时间、地点和关键信息（费用、时长、亮点等）[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=11%3A30%E2%80%9312%3A00%20%7C%20Check,thrilling%20attractions%20such%20as%20Volcano)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=18%3A00%E2%80%9319%3A30%20,exhibits%2C%20and%20animal%20behavior%20demonstrations)。针对示例查询，上述行程涵盖了用户偏好的景点和舒适的交通安排。

评估方面，TripTailor将自动检查生成行程的合规性和匹配度。系统根据可行性指标验证是否遗漏重要安排，依据合理性指标判断行程是否高效、费用是否合规，依据个性化指标判断行程是否体现用户兴趣[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。生成行程最终被打分或标记优劣，研究者可对比不同模型输出的评分，分析其优势和缺陷[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)。

**实验亮点与主要发现**

* **LLM表现不足：** 在TripTailor测试中，绝大多数LLM生成行程未达到人类规划水平。实验显示只有不到10%的LLM行程在综合质量上超越了真实行程[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=020%20providing%20a%20more%20authentic,agents%20capable%20of%20understanding%20and)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。
* **可行性和合理性问题：** 许多生成行程未通过基本可行性检查（如遗漏住宿或交通信息），或包含不合理路线（同日景点间距离过大）。即使在最优策略下，可行率和合理率也未达到100%，多日行程的路线优化依然是挑战[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Evaluation%20Metrics%20Description%20Feasibility%20Within,not%20be%20repeated%20throughout%20the)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。
* **个性化挑战：** 当前模型难以充分满足用户偏好。绝大多数LLM输出在个性化评估中不及真实行程；只有极少数案例的生成行程被评为“更符合用户偏好”[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。这表明模型在理解和执行用户高层需求（如主题偏好、活动类型）方面还有待提高[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Personalization%20Individual%20Preference%20The%20plan,habits%2C%20and%20other%20relevant%20factors)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。
* **路线效率不足：** 统计显示LLM生成行程的平均旅行距离通常显著大于真实行程（即使最优策略平均距离也达真实的1.8倍，一般方法更高），说明模型未能有效优化行程顺序[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=recommended%20time%20range,habits%2C%20and%20other%20relevant%20factors)。
* **启发与挑战：** 这些结果提示，未来需改进模型对多约束规划的理解和对用户偏好的捕捉，以及更有效地使用检索与验证工具[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=5%20Conclusion%20540%20We%20introduce,548)[openreview.net](https://openreview.net/pdf?id=cdggAAy8DK#:~:text=Average%20Route%20Feasibility%20Rationality%20Personalization,3)。

**📌 TripTailor 技术报告（用于分享展示）**

**1️⃣ 项目背景与研究意义**

随着大型语言模型（LLM）推理和组合能力不断增强，自动化旅行规划逐渐成为AI Agent领域的重要应用场景。然而，现有旅行规划工具大多依赖人工模板或简单的规则组合，缺乏真正可落地的全流程自动化与多维度评测。现有基准如 **TravelPlanner**、**ChinaTravel**，要么依赖模拟数据、要么规模有限，难以真实还原复杂旅行场景和个性化需求。因此，TripTailor 旨在构建一个面向真实用户需求、规模化、全流程可复现的旅行规划与评测基准，推动 LLM 在复杂现实任务中的落地研究。

**2️⃣ 数据集规模与内容**

TripTailor 核心由三部分组成：

* **兴趣点库（POI Sandbox）**：涵盖 40 个中国热门城市，约 5.6k 景点、8.9 万酒店、42 万餐厅、1.5 万航班与 2.8 万火车路线，均含价格、位置、评分、类别、标签、开放时间等丰富元信息。
* **真实行程对**：近 4000 条真实采集和重写后的多日行程，包含每天详细时间线、交通、住宿、餐饮和景点安排。
* **用户查询**：从行程中自动抽取关键信息，由 LLM 生成自然语言查询，覆盖预算、天数、交通偏好、活动偏好、餐饮偏好等。

**3️⃣ 系统使用流程示例（文字演示）**

**示例：**  
用户提出请求：

“我想从成都去上海，玩 2 天，周四早晨出发周五晚上回，预算 3100 元，住中档 4 星酒店，每餐 50-100 元，想去海洋馆、动物园，既想参观又要有互动体验，行程不要太紧张。”

**TripTailor 系统执行：**

1. **交通工具检索**：航班查询工具从沙盒中找到 07:45 出发航班 CA4503，10:35 到达上海。
2. **住宿检索**：酒店查询工具匹配“4 星中档”，选出如上海文飞精品酒店等候选项。
3. **景点匹配**：景点搜索工具过滤出如上海海昌海洋公园、野生动物园等匹配“海洋 & 野生动物”主题的 4A 以上景点。
4. **餐厅匹配**：餐厅查询工具在景点附近筛选 50~100 元区间内的餐厅（如“在水一方·鱼上世界”火锅）。
5. **行程生成**：LLM 将以上信息拼装成时间线：
   * 07:45-10:35 航班抵达
   * 11:30 入住酒店
   * 13:00-17:00 海洋公园
   * 18:00 晚餐
   * 第二天上午野生动物园 → 午餐 → 下午前往机场，16:25 返回成都。
6. **自动化评估**：TripTailor 核对是否超预算、交通安排是否匹配沙盒、景点餐饮是否合理多样，生成可行性与个性化评分报告。

**4️⃣ 技术实现过程与关键方案**

**（1）沙盒环境构建**：

* 数据采集自公开 OTA 与高德地图，POI 自动补全经纬度，筛选真实用户高评分景点。
* 所有信息在规划环节均来自沙盒，不允许模型自造“幻觉 POI”。

**（2）多工具链 + LLM 协作**：

* 设计独立工具用于交通、住宿、景点、餐饮检索；LLM 仅负责调度、组合、时间线生成。
* 所有信息可溯源，方便评测与复现。

**（3）个性化查询自动生成**：

* 提示 LLM 严格使用“第一人称”生成查询，不包含具体地点名，仅抽象描述偏好。
* 保证与真实行程一一对应，可用于对比和训练奖励模型。

**（4）高可信度自动评估**：

* 可行性：检查是否漏交通/住宿、是否出现沙盒外信息。
* 合理性：检查餐饮/景点重复、路线是否最优、预算是否超出。
* 个性化：LLM + 奖励模型双重比对真实行程，看是否更符合偏好。

**5️⃣ Workflow 方法核心流程**

TripTailor 提出了一个可复制的**Workflow 分解流程**，模拟人类旅行策划的思维顺序：  
**步骤 1️⃣：交通决策**  
先锁定出发和返程交通（航班/火车），保证时间可行。  
**步骤 2️⃣：核心景点优选**  
根据用户主题偏好从 POI 库筛选候选景点 → LLM 排序 → 选定每日游览核心点。  
**步骤 3️⃣：住宿选址**  
根据景点分布确定酒店位置（中心/靠近某景点），确保路线紧凑。  
**步骤 4️⃣：餐饮补充**  
为每日景点匹配附近餐厅，优先选符合用户餐标和口味的组合。  
**步骤 5️⃣：时间线生成**  
由 LLM 按顺序编排，保证不冲突、时间衔接合理。  
**步骤 6️⃣：自动化检查与微调**  
最后对比预算、检查幻觉、检查合理性，必要时回溯步骤调整。

该工作流显著提升了行程可执行性与合理性，在基准实验中相比直接生成有较大提升，但在“个性化细节”方面仍存在瓶颈，表明未来需要更强的偏好对齐与上下文记忆能力。

**6️⃣ 已解决的典型问题与不足**

✅ **已解决：**

* 提供了可控沙盒，杜绝 LLM 编造信息。
* 将规划拆解为多个可解释步骤，显著提升成功率。
* 提供真实人类行程对比，科学衡量个性化水平。

❌ **仍待突破：**

* LLM 在多日行程路线优化上仍存在无效绕路问题。
* 个性化超越率仍低于人类，尤其在主题深度与活动设计上。
* 当前只能生成单轮结果，缺少真实场景中常见的“多轮对话”反复调整。

**📌 总结**

TripTailor 以真实数据、可复现的工作流和严格的三维度评估，成为 LLM 旅行规划落地研究的可用标准。其开源工具和基线方法可帮助研究者快速验证新模型和策略，是探索 Agent 多步骤推理、规划可解释性和用户偏好对齐的重要测试场景。