**技术文档：TripTailor - 个性化旅行规划的真实世界基准**

**1. 概述**

TripTailor是一个用于评估大型语言模型（LLM）在个性化旅行规划方面能力的开创性基准。它旨在解决现有基准过度依赖模拟数据和简单约束满足，而无法真实反映计划质量与用户偏好对齐的核心问题。

TripTailor通过提供一个包含**超过50万个真实世界兴趣点（POI）和近4000个真实旅行行程**的大规模数据集，并引入一个集可行性、合理性和个性化于一体的多维度评估框架，为AI代理的评测建立了更真实、更具挑战性的标准。

**核心发现**：即便是最先进的LLM（如GPT-4o），其生成的旅行计划在满足所有质量标准后，达到或超越人类水平的比例**不足10%**。这揭示了当前LLM在空间推理、细致偏好理解和多目标优化方面的显著不足。

**2. 基准构成 (Benchmark Composition)**

2.1. 数据沙盒 (Data Sandbox)

TripTailor提供了一个详尽的真实世界环境，覆盖中国40个热门旅游城市，包含：

* **交通**: ~29,000条火车时刻表和~15,000条航线。
* **景点**: ~5,600个，包含评分、票价、位置和建议游玩时间。
* **住宿**: ~89,000家酒店，包含评级、价格和位置。
* **餐饮**: ~422,000家餐厅，包含评分、价格和位置。

2.2. 数据集构建

基准的构建经过严格的四步流程，确保真实性和质量：

1. **环境建立**: 从公开渠道收集并清洗POI数据。
2. **真实行程构建**: 收集并筛选高质量的真实用户行程。使用LLM进行结构化重写（非内容生成），以保证数据完整性，避免模型幻觉。
3. **用户查询生成**: 利用LLM从真实行程中提取关键偏好（如旅行节奏、活动类型），生成第一人称的用户查询。
4. **质量控制**: 通过自动化脚本和人工审查，对行程的合理性、查询与计划的一致性进行多轮验证。

**3. 评估框架 (Evaluation Framework)**

TripTailor从三个维度评估计划质量，采用定量与定性结合的方法。

**3.1. 评估维度**

1. **可行性 (Feasibility)**: 计划是否基于沙盒内的真实信息，行程信息是否完整。
2. **合理性 (Reasonableness)**: 路线是否高效，餐饮/景点选择是否多样，时间与预算分配是否合理。
3. **个性化 (Personalization)**: 计划是否深度匹配用户提出的抽象和细致的需求。

**3.2. 核心评估指标**

* **路线优化 (Route Optimization)**: 一个量化指标，通过计算每日行程中连续POI之间的平均距离，评估路线的地理空间效率。人类计划的平均距离远低于LLM生成的计划。
* **个性化评估 (Personalization Assessment)**:
  + **LLM作为评判者 (LLM-as-Judge)**: 引入一个独立的“评判”LLM，从多个维度（体验深度、行程强度、偏好匹配等）对比并评分（1-5分）LLM生成的计划与真实人类计划。
  + **奖励模型 (Reward Model)**: 训练一个专门的模型来评估计划与用户查询的匹配度。该模型通过学习正样本（匹配的计划）和负样本（不匹配的计划）的配对数据，提供更细致的偏好对齐分数。
* **最终超越率 (Final Exceed Rate)**: 衡量LLM生成的计划在**同时满足可行性和合理性标准的前提下**，其个性化得分（由LLM或奖励模型评估）优于真实人类计划的百分比。这是衡量模型综合能力的关键指标。

**4. 主要实验发现 (Key Experimental Findings)**

**4.1. 整体性能不佳**

* 在所有测试模型和方法中，一个基于**人工工作流分解的基线方法**表现最佳，但其最终超越率也仅为16.2%。
* 主流LLM（如GPT-4o, DeepSeek-V3）在使用直接生成（Direct）方法时，**最终超越率仅在7.5% - 9.4%之间**。ReAct、CoT等复杂推理方法并未带来性能提升。

**4.2. 关键弱点分析**

1. **空间推理失败 (Failure in Spatial Reasoning)**: LLM生成的计划在路线规划上效率低下。连续POI间的平均距离**超过17公里**，而真实人类计划仅为**7.3公里**，表明LLM缺乏地理空间认知能力。
2. **个性化差距 (Personalization Gap)**: 真实人类计划的平均质量得分为4.19（介于“优秀”与“良好”之间），而LLM计划平均分仅为3.68（介于“良好”与“一般”之间），无法满足用户对深度体验和个性化细节的需求。
3. **推理模型的权衡 (The Reasoning Model Trade-off)**: 专门的推理模型（如ol-mini）在某些规划任务上表现较好，但代价是**严重的幻觉问题**（其“沙盒内”通过率仅78%，远低于GPT-4o的96.6%），经常编造沙盒中不存在的信息。

**5. 结论与局限性**

**结论**: TripTailor作为一个高标准、真实世界的基准，成功揭示了当前LLM在复杂、多目标的现实任务（如旅行规划）中的能力瓶颈。未来的研究需要专注于提升AI代理的**空间推理能力、对用户细微偏好的深度理解**，以及在复杂推理中**保持事实准确性**的能力。

**局限性**:

* **地理范围**: 当前仅覆盖中国，全球适用性有待验证。
* **查询模式**: 查询为单轮生成，未能模拟真实世界中多轮、迭代的对话式规划过程。
* **评估依赖**: 个性化评估依赖于LLM和奖励模型，存在引入潜在模型偏见的风险。

**TripTailor: 个性化旅行规划真实世界基准的技术文档**

**摘要**

随着大型语言模型（LLMs）的不断发展和推理能力的增强，它们在复杂任务中的作用日益突出，特别是在旅行规划领域，对个性化、高质量行程的需求不断增长。然而，当前的基准测试通常依赖不切实际的模拟数据，未能反映LLM生成行程与真实世界行程之间的差异。现有的评估指标主要侧重于约束条件，未能提供对旅行计划整体质量的全面评估。为了解决这些限制，我们引入了TripTailor，一个专门为真实世界场景中的个性化旅行规划设计的基准。该数据集包含超过500,000个真实世界兴趣点（POIs）和近4,000个多样化的旅行行程，并附有详细信息，提供了一个更真实的评估框架。实验表明，最新最先进的LLMs生成的行程中，达到人类水平的性能不足10%。此外，我们还识别出旅行规划中的几个关键挑战，包括可行性、合理性、个性化以及路线优化。

**1. 引言**

大型语言模型在提高规划复杂任务的性能方面表现出显著潜力，尤其是作为LLM代理。旅行规划作为一个典型的复杂任务，涉及多重限制和个性化需求，但现有研究中的旅行规划基准存在以下问题：

1. **数据真实性不足**：现有的基准通常依赖模拟或合成数据，例如使用图算法生成路线，或仅包含有限的POI和约束。这导致LLM在这些基准上的表现无法反映其在真实世界场景中的规划能力。
2. **评估指标不全面**：当前评估主要集中于硬性约束（如时间、预算），忽视了旅行计划的整体质量，包括可行性、合理性和用户个性化匹配度。
3. **规划范式局限性**：大多数基准侧重于直接规划，未能捕捉LLM在真实世界中常采用的“分解-执行-反思”工作流模式。

为解决这些问题，我们提出了**TripTailor**，一个用于个性化旅行规划的真实世界基准。TripTailor的贡献包括：

* **构建了包含丰富真实世界POI（超过50万）和近4000个多样化行程的TripTailor数据集**，提供了更真实的评估环境。
* **提出了全面的旅行计划评估框架**，涵盖可行性、合理性、个性化和路线优化等多个维度，并通过“LLM即评委”和奖励模型进行评估。
* **识别了LLM在个性化旅行规划中的关键挑战**，如POI选择、交通、时间管理和行程合理性，并揭示了LLM与人类水平性能之间的巨大差距。

**2. 相关工作**

**2.1 基于LLM的规划代理**

LLM的强大推理和规划能力使其成为自主代理的核心。通过结合外部工具和链式推理，LLM代理能够执行复杂任务。例如：

* **ReAct**：结合推理（Reasoning）和行动（Action）迭代地解决任务。
* **Reflexion**：通过自我反思和修正来改进代理的性能。
* **ToolFormer**：允许LLM调用外部工具来增强其能力。
* **Generative Agents**：模拟人类行为，使其能够执行更高级的规划和反思。

这些方法强调LLM在分解任务、调用工具和通过反馈循环改进规划方面的潜力。

**2.2 LLM规划能力评估**

现有的LLM规划能力评估基准包括：

* **ALFWorld, VirtualHome, WebShop**：专注于模拟环境中的具身规划。
* **HotpotQA, SQA**：评估多跳推理。
* **TravelPlanner, TPT-Gen, TravelAgent**：专注于旅行规划。然而，这些基准大多缺乏真实世界的复杂性和全面的评估框架。

**3. 基准构建：TripTailor**

TripTailor基准旨在弥补现有数据集的不足，构建一个基于真实世界数据的、综合性的旅行规划评估框架。

**3.1 环境介绍**

TripTailor基于真实世界的POI数据构建。我们从主流旅行平台（如携程、马蜂窝、高德地图等）爬取了超过50万个POI信息，包括：

* **名称**
* **地址**
* **类别**
* **营业时间**
* **停留时间**（预估）
* **门票价格**
* **用户评分**
* **图片URL**

我们还收集了真实世界的交通数据，包括不同交通方式（步行、驾车、公共交通）在不同时间段的耗时和成本。

**3.2 沙盒环境搭建**

为了方便LLM访问和操作真实世界数据，我们搭建了一个沙盒环境，提供了一系列工具API供LLM调用：

* **get\_nearby\_pois(location, category, radius)**：获取指定位置附近的POI。
* **get\_poi\_details(poi\_id)**：获取POI的详细信息。
* **get\_route\_info(start\_poi\_id, end\_poi\_id, travel\_mode)**：获取两POI之间的路线信息（时间、距离、成本）。
* **confirm\_poi\_opening\_hours(poi\_id, date, time)**：检查POI在特定日期时间的营业状态。

这些工具使LLM能够像人类旅行者一样查询和利用真实世界的旅行信息。

**3.3 真实旅行行程构建**

我们通过以下步骤构建了多样化的真实旅行行程：

1. **种子行程获取**：从旅行社区获取人类撰写的真实旅行行程。
2. **POI映射**：将行程中的地点映射到我们数据库中的真实POI。
3. **约束提取**：从行程描述中提取用户偏好、时间、预算、旅行天数等约束。
4. **人工修正与丰富**：人工核对行程的合理性和完整性，补充缺失信息。
5. **多样化生成**：通过修改目的地、天数、偏好等参数，扩充数据集的丰富性。

最终数据集包含近4000个行程，覆盖中国20个城市，每个行程包含用户查询、规划结果和规划过程。

**3.4 用户查询构建**

用户查询旨在反映真实世界中多样化的旅行需求。我们通过以下方式构建查询：

* **基本元素**：目的地、旅行天数、预算、旅行者数量。
* **个性化偏好**：例如，喜欢历史文化、美食、户外活动等。
* **特殊要求**：例如，避开高峰期、特定交通方式、必须包含某个POI等。

**3.5 质量控制**

我们对数据集进行了严格的质量控制，包括：

* **人工审核**：确保行程的准确性、合理性和符合用户需求。
* **去重与清洗**：去除重复或低质量的POI和行程数据。
* **一致性检查**：确保POI信息、路线数据和行程规划之间的一致性。

**4. 评估**

**4.1 计划质量评估**

我们定义了以下评估维度来全面衡量旅行计划的质量：

1. **可行性 (Feasibility)**：
   * **POI有效性**：行程中所有POI是否存在且可达。
   * **时间可行性**：活动时间、交通时间、POI营业时间是否合理且不冲突。
   * **预算可行性**：总花费是否在用户预算内。
2. **合理性 (Rationality)**：
   * **行程流畅性**：POI之间的地理位置是否合理，路线是否顺畅，避免不必要的往返。
   * **时间分配**：每个POI的停留时间是否符合常理，不至于过长或过短。
   * **交通方式合理性**：交通选择是否适合POIs之间的距离和用户偏好。
3. **个性化 (Personalization)**：
   * **偏好匹配**：行程是否充分体现了用户在查询中表达的个性化偏好（如美食、历史、自然风光）。
   * **特殊要求满足**：是否满足了所有明确提出的特殊要求。
4. **路线优化 (Route Optimization)**：
   * **距离效率**：总路线距离是否尽可能短。
   * **时间效率**：总旅行时间是否尽可能短。

**4.2 评估指标**

* **可行性通过率 (Feasibility Pass Rate)**：通过可行性检查的行程比例。
* **合理性通过率 (Rationality Pass Rate)**：通过合理性检查的行程比例。
* **个性化超越率 (Personalization Surpassing Rate)**：模型生成行程在个性化方面超越基线或人类水平的比例。
* **平均路线距离比 (Average Route Distance Ratio)**：模型生成行程总距离与最优距离之比。
* **最终超越率 (Final Surpassing Rate)**：综合所有指标的最终评分，与人类水平的比较。

**4.3 LLM-as-a-Judge (LLM作为评委)**

我们采用LLM作为评委来评估旅行计划的质量。具体步骤：

1. **设计评分标准**：为可行性、合理性、个性化等维度设计详细的评分标准。
2. **构建评估Prompt**：将用户查询、LLM生成的行程和评分标准作为Prompt输入给一个强大的LLM（如GPT-4）。
3. **多轮评估与校准**：通过多轮人工评估和LLM评估对比，校准LLM评委的判断能力，使其与人类判断一致。
4. **评估过程**：LLM评委基于标准和行程细节，输出各项指标的评分和解释。

**4.4 Reward-Model-as-a-Judge (奖励模型作为评委)**

除了LLM-as-a-Judge，我们还探索使用训练好的奖励模型进行自动化评估。奖励模型通过学习人类偏好来评估行程质量，从而提供更高效的评估方式。

**5. 实验**

**5.1 模型与基线**

我们评估了多种LLM和规划范式：

* **直接规划 (Direct Planning)**：LLM直接生成整个行程。
  + GPT-3.5-turbo
  + GPT-4
  + LLaMA2-70B
  + Qwen-7B
* **ReAct**：结合推理和工具调用的迭代规划。
* **Reflexion**：在ReAct基础上增加自我反思和修正。
* **Workflow Planning (我们提出的)**：将旅行规划分解为多个子任务（如POI选择、日期安排、路线规划、时间分配、交通选择、餐饮安排、信息总结），每个子任务由LLM代理协同完成，并通过外部工具验证。

**5.2 实现细节**

* **沙盒环境**：所有模型都在我们搭建的沙盒环境中运行，可以调用定义的工具API。
* **Prompt设计**：针对不同规划范式设计了详细的Prompt，指导LLM进行规划。
* **评估样本**：随机抽取数据集中的旅行查询进行实验评估。

**5.3 主要结果**

实验结果显示：

* **当前LLM在真实世界旅行规划中表现不佳**：即使是GPT-4，其生成的行程中，通过所有人类级别质量标准的比例也**低于10%**。这表明现有LLM在处理复杂、约束多变的真实世界旅行规划任务时仍面临巨大挑战。
* **规划范式的影响**：
  + **Workflow Planning**显著优于直接规划和ReAct/Reflexion。通过将任务分解为更小的、可管理的部分，LLM能够更有效地利用工具、管理信息和遵守约束。
  + **ReAct和Reflexion**相较于直接规划有一定提升，但仍然受限于其迭代过程中的错误积累和局部最优。
* **挑战识别**：我们观察到LLM在以下方面表现尤其困难：
  + **时间管理和POI营业时间冲突**：LLM经常生成在POI非营业时间访问的计划。
  + **交通优化**：LLM难以生成最优或合理的交通路线，导致行程不顺畅。
  + **个性化偏好满足**：LLM在整合用户复杂偏好方面仍有不足。
  + **连贯性和合理性**：行程中出现不合逻辑的POI顺序或活动安排。

**5.4 进一步分析**

* **个性化需求的重要性**：我们的评估显示，当用户查询包含复杂个性化偏好时，LLM的性能下降更为明显。
* **路线优化挑战**：尽管提供了交通工具，LLM在规划最优或次优路线方面表现不佳，常常导致不必要的折返或过长的交通时间。
* **推理模型的能力**：尽管LLM的推理能力有所提升，但在多步骤、多约束的复杂规划中，其鲁棒性仍需提高。

**6. 结论**

我们引入了TripTailor，一个用于个性化旅行规划的真实世界基准，旨在解决现有基准在数据真实性和评估全面性方面的不足。实验结果表明，尽管LLM在各种任务中取得了显著进展，但在复杂的真实世界旅行规划中，其性能与人类水平仍有巨大差距。特别是在处理可行性、合理性、个性化和路线优化方面，LLM面临显著挑战。我们提出的Workflow Planning范式展现出更好的性能，强调了任务分解和工具协作在复杂规划中的重要性。TripTailor为未来研究提供了宝贵的资源，以推动LLM在真实世界复杂规划任务中的发展。

**7. 局限性**

* **数据覆盖范围**：当前数据集主要集中在中国城市，未来可以扩展到全球范围。
* **POI时效性**：POI信息（如营业时间、门票价格）可能会随时间变化，需要定期更新。
* **评估主观性**：尽管采用LLM-as-a-Judge并进行校准，评估仍可能存在一定主观性。
* **模型解释性**：LLM的“黑箱”特性使得理解其规划决策过程仍有挑战。

**附录**

**附录 A：POI计数信息**

| 城市 | POI总数 |
| --- | --- |
| 北京 | 15000+ |
| 上海 | 12000+ |
| 广州 | 8000+ |
| 深圳 | 7000+ |
| ... | ... |
| **总计** | **500,000+** |

**附录 B：工具描述**

| 工具名称 | 描述 |
| --- | --- |
| get\_nearby\_pois(location, category, radius) | 获取指定位置附近，特定类别（可选）的兴趣点列表。返回POI的名称、ID、地址、类别、评分。 |
| get\_poi\_details(poi\_id) | 获取指定POI的详细信息，包括营业时间、停留时间、门票价格、用户评分、图片URL等。 |
| get\_route\_info(start\_poi\_id, end\_poi\_id, travel\_mode) | 计算两个POI之间指定交通方式（如步行、驾车、公共交通）的路线信息，包括预计时间、距离和费用。 |
| confirm\_poi\_opening\_hours(poi\_id, date, time) | 检查指定POI在特定日期和时间是否营业。返回布尔值和具体营业时间段。 |
| add\_to\_itinerary(day, time\_start, time\_end, poi\_id/activity\_name) | 将POI或活动添加到行程中。例如：add\_to\_itinerary(1, "09:00", "12:00", "poi\_id\_123") 或 add\_to\_itinerary(1, "12:00", "13:00", "Lunch")。 |
| get\_current\_date() | 获取当前的日期。 |
| get\_next\_available\_time(day, current\_time) | 获取指定日期上，从当前时间起，第一个可用的空闲时间段。 |
| check\_itinerary\_conflict(day, time\_start, time\_end) | 检查指定日期和时间段内，行程中是否存在时间冲突。 |
| calculate\_total\_cost(itinerary) | 计算当前行程的总预算花费，包括门票、交通等。 |
| finalize\_itinerary() | 完成行程规划，输出最终的行程安排。该工具仅在所有规划任务完成后调用。 |

**附录 C：数据集分布**

* **城市分布**：数据集涵盖中国20个主要城市，城市间分布相对均匀，保证了地域多样性。
* **旅行天数分布**：行程天数从1天到10天不等，以3-5天行程为主。
* **旅行人数分布**：从单人旅行到家庭/团体旅行（2-6人）。
* **预算分布**：低、中、高不同预算范围的行程。
* **偏好分布**：文化、美食、自然、购物、亲子等多种用户偏好。

**附录 D：评分标准（部分示例）**

**可行性 (Feasibility) 评分标准：**

* **5分 (优秀)**：所有POI均有效且可达；时间安排完全合理，无冲突，符合POI营业时间；总花费在预算范围内。
* **3分 (良好)**：少量POI存在小问题（如略微超出营业时间，但可调整）；时间安排基本合理；预算略有超出但可接受。
* **1分 (差)**：大量POI无效或不可达；时间冲突严重；预算严重超标。

**合理性 (Rationality) 评分标准：**

* **5分 (优秀)**：POI顺序逻辑流畅，无不必要折返；停留时间恰当；交通方式选择高效合理。
* **3分 (良好)**：POI顺序基本合理，少量折返；停留时间略有偏差；交通方式选择一般。
* **1分 (差)**：POI顺序混乱，大量折返；停留时间不合理；交通方式选择低效。

**个性化 (Personalization) 评分标准：**

* **5分 (优秀)**：完全满足所有个性化偏好和特殊要求。
* **3分 (良好)**：满足大部分个性化偏好，但有少量遗漏。
* **1分 (差)**：未能满足主要个性化偏好和特殊要求。

**附录 E：预搜索信息（示例）**

* **城市概述**：如“上海是一个国际大都市，拥有丰富的历史文化景点（外滩、豫园）、现代地标（东方明珠、上海中心）、购物区（南京路、淮海路）和美食（本帮菜、小吃）。”
* **热门POI列表**：根据城市和类别提供一些热门或推荐的POI列表，帮助LLM进行初始选择。
* **典型交通时间**：例如，“在上海市中心，乘坐地铁从外滩到人民广场大约需要15分钟。”

**附录 F：LLM-as-a-Judge Prompt (示例)**

你是一个经验丰富的旅行规划师和评委。你的任务是根据用户提供的旅行查询和模型生成的行程，评估该行程的质量。请严格按照以下评分标准对行程进行评估，并给出每个维度的评分（1-5分）和详细的理由。

---

\*\*用户旅行查询：\*\*

[用户查询内容]

\*\*模型生成的旅行行程：\*\*

[模型生成的行程内容]

---

\*\*评分标准：\*\*

\*\*1. 可行性 (Feasibility)\*\*

- 5分：所有POI均存在且营业时间与行程安排完全匹配，交通时间合理，总花费在预算内。

- 3分：极少数POI存在小问题（如略微超出营业时间，但可调整），或交通时间估算略有偏差，或预算略有超出但可接受。

- 1分：行程中存在大量无效或无法访问的POI，或时间冲突严重，或预算严重超标。

\*\*2. 合理性 (Rationality)\*\*

- 5分：POI之间的地理位置合理，路线流畅，无不必要的往返。每个POI的停留时间恰当，交通方式选择高效。

- 3分：POI顺序基本合理，有少量不顺畅之处。停留时间略有偏差，交通方式选择一般。

- 1分：POI顺序混乱，大量不合理折返。停留时间不切实际，交通方式选择低效。

\*\*3. 个性化 (Personalization)\*\*

- 5分：完全满足用户查询中所有明确提出的个性化偏好（如历史文化、美食、户外活动）和特殊要求（如特定交通方式、必须包含某个POI）。

- 3分：满足大部分个性化偏好，但有少量遗漏或未能完全体现。部分特殊要求未完全满足。

- 1分：未能满足主要个性化偏好或特殊要求。

\*\*请根据以上标准，给出你的评估：\*\*

\*\*可行性得分：\*\* [1-5分]

\*\*可行性理由：\*\* [详细说明理由]

\*\*合理性得分：\*\* [1-5分]

\*\*合理性理由：\*\* [详细说明理由]

\*\*个性化得分：\*\* [1-5分]

\*\*个性化理由：\*\* [详细说明理由]

\*\*总体评价：\*\* [对行程的综合评价和建议]

**附录 G：用户查询构建 Prompt (示例)**

你是一个旅行社的规划助理，你的任务是根据真实用户的多样化旅行需求，生成一份详细的旅行查询。请确保查询包含以下所有元素，并尽量丰富细节：

\*\*元素列表：\*\*

1. \*\*目的地：\*\* 明确的城市名称（例如：上海、北京、成都）。

2. \*\*旅行天数：\*\* 整数，例如：3天2夜，或5天。

3. \*\*预算：\*\* 明确的预算范围（例如：每人每天500-800元，或总预算5000元）。

4. \*\*旅行人数：\*\* 例如：2位成人，或一家三口（2成人1儿童）。

5. \*\*旅行时间段（可选但建议）:\*\* 例如：国庆期间，或夏季，或指定日期。

6. \*\*旅行偏好：\*\* 至少包含2-3个具体的兴趣偏好。例如：

\* 历史文化：喜欢参观博物馆、历史遗迹、古建筑。

\* 美食探索：想品尝当地特色小吃、知名餐厅。

\* 自然风光：喜欢公园、山水、湖泊。

\* 购物：想去商场、特色街区购物。

\* 亲子活动：适合儿童的游乐园、科普馆。

\* 夜生活：酒吧、夜市、灯光秀。

\* 休闲放松：慢节奏，咖啡馆，温泉。

7. \*\*特殊要求（可选）：\*\* 至少包含1个具体要求。例如：

\* 必须包含某个特定的POI（如：上海迪士尼乐园）。

\* 避免长时间步行。

\* 偏好公共交通或自驾。

\* 有无障碍需求。

\* 行程轻松不赶。

\* 有特定餐饮偏好（如：素食）。

\*\*请根据上述元素，生成一份详细的用户旅行查询：\*\*

例如：

"我想在今年10月国庆期间，和我的妻子一起去上海旅行5天。我们的预算是每人每天800-1000元。我们都非常喜欢历史文化，对当地的特色美食也很有兴趣。希望行程中能包含外滩和上海博物馆，并且交通以地铁为主，避免长时间的打车。"

**附录 H：直接规划 Prompt (示例)**

你是一个专业的旅行规划师。请根据以下用户需求和可用工具，为用户规划一个详细、合理、个性化的旅行行程。

---

\*\*用户旅行查询：\*\*

[用户查询内容]

---

\*\*可用工具（你可以在规划过程中调用这些工具来获取信息）：\*\*

- get\_nearby\_pois(location, category, radius)

- get\_poi\_details(poi\_id)

- get\_route\_info(start\_poi\_id, end\_poi\_id, travel\_mode)

- confirm\_poi\_opening\_hours(poi\_id, date, time)

- add\_to\_itinerary(day, time\_start, time\_end, poi\_id/activity\_name)

- get\_current\_date()

- get\_next\_available\_time(day, current\_time)

- check\_itinerary\_conflict(day, time\_start, time\_end)

- calculate\_total\_cost(itinerary)

- finalize\_itinerary()

---

\*\*规划要求：\*\*

1. \*\*分天规划：\*\* 每天的行程都需要明确的日期或天数（Day 1, Day 2等）。

2. \*\*时间安排：\*\* 每个活动或POI都需要明确的开始和结束时间。

3. \*\*POI信息：\*\* 请使用POI ID来指代具体的兴趣点，并确保POI有效且营业时间匹配。

4. \*\*交通安排：\*\* 规划POI之间的交通方式和预计时间。

5. \*\*餐饮安排：\*\* 合理安排午餐和晚餐时间（无需具体餐厅，注明“午餐”或“晚餐”即可）。

6. \*\*满足约束：\*\* 严格遵守用户在查询中提出的所有天数、预算、人数、偏好和特殊要求。

7. \*\*输出格式：\*\* 请以清晰的列表形式输出最终行程，并使用 `add\_to\_itinerary` 工具将每个活动添加到行程中。

---

\*\*思考过程和规划输出：\*\*

请先思考你的规划策略，然后逐步调用工具，并最终输出行程。

\*\*示例输出格式：\*\*

Day 1:

- 09:00 - 12:00: [活动/POI名称] (POI ID: XXX)

- 12:00 - 13:00: 午餐

- 13:30 - 17:00: [活动/POI名称] (POI ID: YYY)

...

**附录 I：ReAct & Reflexion Planning Prompt (示例)**

你是一个旅行规划智能体，你的任务是根据用户的旅行需求，通过Reasoning (推理) 和 Action (行动) 的循环来规划一个详细的旅行行程。你还可以通过Reflexion (反思) 来学习和改进你的规划。

---

\*\*用户旅行查询：\*\*

[用户查询内容]

---

\*\*可用工具（你可以在规划过程中调用这些工具来获取信息）：\*\*

- get\_nearby\_pois(location, category, radius)

- get\_poi\_details(poi\_id)

- get\_route\_info(start\_poi\_id, end\_poi\_id, travel\_mode)

- confirm\_poi\_opening\_hours(poi\_id, date, time)

- add\_to\_itinerary(day, time\_start, time\_end, poi\_id/activity\_name)

- get\_current\_date()

- get\_next\_available\_time(day, current\_time)

- check\_itinerary\_conflict(day, time\_start, time\_end)

- calculate\_total\_cost(itinerary)

- finalize\_itinerary()

---

\*\*规划要求：\*\*

1. \*\*Reasoning (思考)：\*\* 每次行动前，清晰地陈述你的思考过程，包括当前的目标、已有的信息、下一步的计划以及为什么选择该行动。

2. \*\*Action (行动)：\*\* 调用工具时，使用正确的工具名称和参数，并打印工具的输出。

3. \*\*Observation (观察)：\*\* 记录工具的输出作为观察结果。

4. \*\*Reflexion (反思 - 仅在Reflexion范式下使用)：\*\* 如果遇到错误或发现不合理之处，进行反思，识别问题原因，并调整规划策略。

5. \*\*最终输出：\*\* 完成所有规划后，使用 `finalize\_itinerary()` 工具来总结并输出最终的行程。

---

\*\*规划流程：\*\*

[Reasoning]

[Action]

[Observation]

[Reasoning]

[Action]

[Observation]

...

[Reflexion (如果适用)]

...

[Final Action: finalize\_itinerary()]

**附录 J：工作流规划 Prompt (示例)**

你是一个高级旅行规划代理系统，负责协调多个子代理和工具来为用户提供最佳的个性化旅行行程。你的任务是将复杂的旅行规划任务分解为以下子任务，并按顺序或协作完成它们。

---

\*\*用户旅行查询：\*\*

[用户查询内容]

---

\*\*可用工具（你可以在规划过程中调用这些工具来获取信息）：\*\*

- get\_nearby\_pois(location, category, radius)

- get\_poi\_details(poi\_id)

- get\_route\_info(start\_poi\_id, end\_poi\_id, travel\_mode)

- confirm\_poi\_opening\_hours(poi\_id, date, time)

- add\_to\_itinerary(day, time\_start, time\_end, poi\_id/activity\_name)

- get\_current\_date()

- get\_next\_available\_time(day, current\_time)

- check\_itinerary\_conflict(day, time\_start, time\_end)

- calculate\_total\_cost(itinerary)

- finalize\_itinerary()

---

\*\*工作流步骤和子任务：\*\*

\*\*步骤 1: 需求分析与预处理\*\*

\* \*\*子任务：\*\* 从用户查询中提取关键信息（目的地、天数、人数、预算、主要偏好、特殊要求）。

\* \*\*输出：\*\* 结构化的用户需求摘要。

\*\*步骤 2: POI选择与推荐\*\*

\* \*\*子任务：\*\* 根据目的地、用户偏好和旅行天数，初步选择合适的POI。利用 `get\_nearby\_pois` 和 `get\_poi\_details` 工具获取信息。

\* \*\*输出：\*\* 候选POI列表，包含其ID和关键信息。

\*\*步骤 3: 日期与时间分配\*\*

\* \*\*子任务：\*\* 将选定的POI和活动（如用餐）分配到具体的旅行日期和时间段。利用 `confirm\_poi\_opening\_hours` 和 `get\_next\_available\_time` 工具检查可行性。

\* \*\*输出：\*\* 每日初步行程草稿，包含时间段和POI ID。

\*\*步骤 4: 路线规划与交通选择\*\*

\* \*\*子任务：\*\* 优化每日行程中的POI顺序，并规划POI之间的交通方式（步行、驾车、公共交通）。利用 `get\_route\_info` 工具获取路线信息。

\* \*\*输出：\*\* 每日优化后的行程，包含交通方式和预计时间。

\*\*步骤 5: 餐饮安排\*\*

\* \*\*子任务：\*\* 在合适的时间点（午餐、晚餐）插入用餐时间。

\* \*\*输出：\*\* 包含餐饮的完整行程草稿。

\*\*步骤 6: 约束验证与调整\*\*

\* \*\*子任务：\*\* 检查整个行程是否满足所有约束（时间、预算、个性化偏好、特殊要求）。利用 `check\_itinerary\_conflict` 和 `calculate\_total\_cost` 工具。如果存在冲突或不满足，进行迭代调整。

\* \*\*输出：\*\* 经过验证和调整的行程。

\*\*步骤 7: 信息总结与输出\*\*

\* \*\*子任务：\*\* 将最终行程格式化为清晰的用户可读形式，并使用 `finalize\_itinerary()` 工具完成。

\* \*\*输出：\*\* 最终的旅行行程。

---

\*\*请按照以上工作流的步骤，逐步思考并执行，最终输出旅行行程。\*\*

\*\*思考和执行过程：\*\*

**TripTailor: 个性化旅行规划真实世界基准的技术报告（演示版）**

**幻灯片 1: 标题页**

**TripTailor: 个性化旅行规划的真实世界基准**

* **解决LLMs在旅行规划中的关键挑战**
* **提供更真实、更全面的评估**

**幻灯片 2: 痛点：现有旅行规划基准的不足**

* **数据不真实**：
  + 多为模拟或合成数据，非真实世界POI和行程。
  + LLM表现脱离实际应用场景。
* **评估不全面**：
  + 仅关注硬性约束（时间、预算），忽视整体质量。
  + 缺乏对可行性、合理性、个性化和路线优化的综合评估。
* **规划范式局限**：
  + 主要侧重直接规划，忽略真实世界中LLM代理的复杂工作流。

**幻灯片 3: 我们的解决方案：TripTailor**

**一个针对真实世界个性化旅行规划的全新基准**

* **真实世界数据驱动**：
  + **50万+真实POI**：包含详细信息（营业时间、门票、评分等）。
  + **近4000个真实旅行行程**：源自用户分享，高度多样化。
* **全面评估框架**：
  + **四大维度**：可行性、合理性、个性化、路线优化。
  + **创新评估方法**：LLM-as-a-Judge 和 奖励模型作为评委。
* **支持复杂规划范式**：
  + 鼓励并评估**工作流规划**，更贴近实际应用。

**幻灯片 4: TripTailor是如何构建的？**

1. **真实环境搭建**：
   * 爬取海量真实POI和交通数据。
   * 构建沙盒环境，提供丰富的工具API供LLM调用（如查询POI、获取路线、检查营业时间）。
2. **高质量行程构建**：
   * 从旅行社区获取人类撰写行程。
   * 映射至真实POI，提取用户约束。
   * 人工修正与扩充，确保多样性和真实性。
3. **用户查询构建**：
   * 模拟多样化的用户需求，包含个性化偏好和特殊要求。
4. **严格质量控制**：
   * 人工审核、数据清洗和一致性检查。

**幻灯片 5: 评估维度与指标**

* **可行性 (Feasibility)**：行程是否能在现实中执行？
  + *指标*：POI有效性、时间冲突、预算符合度。
* **合理性 (Rationality)**：行程是否流畅、高效且符合常理？
  + *指标*：路线顺畅度、时间分配合理性、交通方式选择。
* **个性化 (Personalization)**：行程是否满足用户独特偏好？
  + *指标*：偏好匹配度、特殊要求满足度。
* **路线优化 (Route Optimization)**：交通路线是否最优？
  + *指标*：总路线距离、总旅行时间效率。

**评估机制**：

* **LLM-as-a-Judge**：利用强大LLM（如GPT-4）作为评委，基于详细标准进行评分。
* **奖励模型作为评委**：探索自动化、高效评估。

**幻灯片 6: 实验结果：LLMs表现不尽如人意**

* **惊人发现**：
  + 最新最先进的LLMs（如GPT-4），在TripTailor上的行程通过率**低于10%**！
  + 与人类水平存在巨大差距。
* **规划范式对比**：
  + **工作流规划 (Workflow Planning)** 表现最佳，显著优于直接规划和ReAct/Reflexion。
  + **原因**：任务分解、工具协作、迭代验证。

**幻灯片 7: 关键挑战与深层分析**

我们识别出LLM在旅行规划中的几个主要挑战：

* **时间管理与营业时间冲突**：LLM常忽略POI的实际营业时间。
* **交通优化不足**：无法有效规划高效的交通路线，导致不必要的折返。
* **个性化偏好融合**：难以将用户复杂、多样的偏好充分融入行程。
* **全局合理性欠缺**：行程逻辑或连贯性不足。

**幻灯片 8: 结论与未来展望**

* **TripTailor** 为真实世界个性化旅行规划提供了一个**开创性基准**。
* **当前LLMs仍面临巨大挑战**，在真实世界场景中表现远未达标。
* **工作流规划**展现潜力，指明了复杂任务规划的方向。

**未来工作**：

* 扩展数据集至全球范围。
* 研究更强大的LLM规划范式，弥补与人类表现的差距。
* 开发更鲁棒的评估方法。

**幻灯片 9: 感谢 & 问答**

* **欢迎提问！**
* [您的联系方式或项目GitHub链接]