Personal Large Language Model Agents: A Case Study on Tailored Travel Planners

Harmanpreet Singh[*] **Nikhil Verma**[*] **Yixiao Wang** **Manasa Bharadwaj**
Homa Fashandi[†] **Kevin Ferreira** **Chul Lee**

LG Electronics, Toronto AI Lab

{harmanpreet.singh, nikhil.verma, yixiao.wang, manasa.bharadwaj
homa.fashandi, kevin.ferreira, clee.lee}@lge.com

Abstract

Large Language Models (LLMs) have made significant progress, becoming more autonomous and capable of handling real-world tasks through their access to tools, various planning strategies, and memory, referred to as LLM agents. One emerging area of focus is customizing these models to cater to individual user preferences, thereby shaping them into *personal LLM agents*. This work investigates how the user model, which encapsulates user-related information, preferences, and personal concepts, influences an LLM agent's planning and reasoning capabilities. We introduce a personalized version of TravelPlanner, called *TravelPlanner+*, and establish baselines for personal LLM agents. Our evaluation strategy contains an LLM-as-a-Judge component, which provides further in-depth insights into the decision-making process of a personal LLM agent by comparing generic and personal plans. Our findings reveal that while generic plans perform robustly, personal plans show marked improvement in relevance and suitability, with preference rates up to 74.4% on validation and 87.3% on the test set. These results highlight the potential of personal LLM agents to significantly enhance user satisfaction.

1 Introduction

AI agents are computational entities that perceive their surroundings, plan, and take actions using tools to complete a task (Xi et al., 2023). Due to the emergent capabilities of Large Language Models (LLMs), augmenting LLMs with reasoning capabilities and tools enables them to act as AI agents (Mialon et al., 2023), i.e., LLM agents. To provide personalized solutions to users, agents need to know their profiles, personal preferences, concepts and understand their *personal queries*. Personal concepts can be anything specific to the user, such as their pet's name (e.g., Charlie), favorite cuisine (e.g., Italian), or home location (e.g., Seattle). Personal queries involve references to these concepts. For example, if a user wants to travel with Charlie, it implies that the accommodation should be pet-friendly. However, the user may not explicitly specify this constraint in their query; instead, they only refer to their personal concept, i.e., Charlie. We call the encapsulation of the user-related information the *user model*. Personalization or adapt-to-user, as defined by (Tseng et al., 2024), offers users an enhanced experience and improves user satisfaction and retention rates. Current LLM research is surely moving towards Personalization, (Salemi et al., 2023; Li et al., 2024). The user's model is tightly integrated into a personal LLM agent.

In this study, we explore the impact of users' models on agent's decision-making and planning processes to create personalized solutions for users. Different environments exist to evaluate the capabilities of LLM agents (Liu et al., 2023). However, none currently incorporate personal user information. We drew inspiration from the TravelPlanner benchmark (Xie et al., 2024), designed to generate a travel plan based on the user's text-based query. TravelPlanner provides a rich and complex environment to test the efficacy of the LLM agents. Like other agent-based benchmarks, the TravelPlanner benchmark offers a generic environment where no personal information or characteristics are provided for the customer. To investigate the effectiveness of the LLMs as a personal agent, in this study, we provide :

- A personalized version of the TravelPlanner, called TravelPlanner+, with user models and personal queries
- Benchmark performance with closed and open source models, on TravelPlanner+, which incorporates user models during the planning
- An evaluation framework to evaluate the

[*]These authors contributed equally to this work.
[†]Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 486–514
November 12-16, 2024

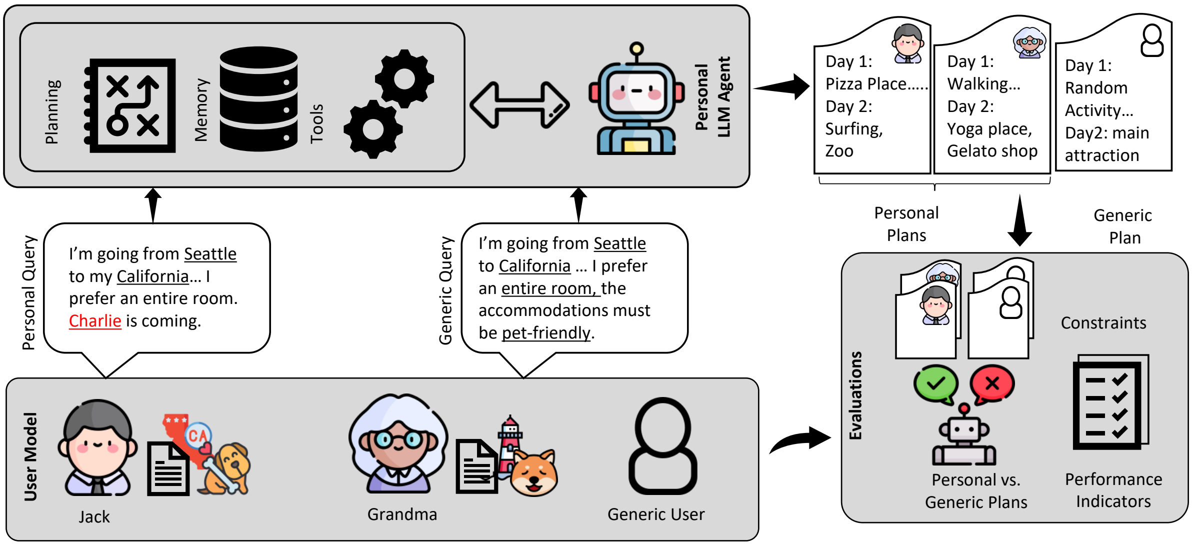


Figure 1: Personal LLM agent understands the customer's profile, preferences, and personal concepts. The customer can communicate with the TravelPlanner+ agent in a customized language. The generated plans are tailored to the customer. Icons by Freepik.

plans generated by personal and generic LLM agents, i.e., LLM-as-a-Judge

Figure 1 shows the overall workflow of the proposed personal LLM agent. The customer is known to the agent regarding their user profile, preferences (e.g., hobbies, cuisine), and personal concepts (e.g., pet name). Due to privacy and security concerns regarding user data, personal LLM agents are better suited to run on devices and provide user data privacy by design. Our focus has been on LLM agents that can be deployed on customers' edge devices. Our workflow, experiments, and evaluation framework offer insight into selecting the optimal model for personal LLM agents and are easily translatable to similar use cases.

2 Background Information

LLM Agents: The agent framework includes the agent, planning, memory, and toolset. LLMs have general-purpose capabilities that make them suitable for use as agents (Wang et al., 2024; Mialon et al., 2023). The planning module helps break down complex tasks into subtasks. It can be implemented through single-path reasoning, such as Chain-of-Thoughts (CoT) (Wei et al., 2022), or multi-path reasoning, such as the Tree-of-Thoughts (Yao et al., 2023a). The planning approaches mentioned so far do not include feedback, making planning for some tasks challenging. Mechanisms such as ReAct (Yao et al., 2023b) and Reflexion (Shinn et al., 2023) allow the model to continuously adjust the execution plan based on past actions and observations. Memory components help maintain past thoughts and interactions. The toolset interacts with the environment to gather detailed information, such as a flight search.

To evaluate the capabilities of LLM-based agents, various benchmarks (Liu et al., 2023) have been developed across different categories, including Code (Zhang et al., 2024; Liao et al., 2023), Game (Hu et al., 2024), and Web (Zhou et al., 2024; Yao et al., 2022; Deng et al., 2024; Xie et al., 2024). We specifically focused on web-based environments and decided to adapt TravelPlanner because it has a wider range of uses for personalization among the general population. Additionally, TravelPlanner presents significant challenges due to its multi-constraints and long-term planning. To make things even more complex, we introduced a personalization element to fully expose and address these challenges.

User Modeling and Personalization: A user model encompasses the data associated with a specific user, including their profile, preferences, and personal concepts (Tan and Jiang, 2023). The user profile includes individual characteristics such as age, gender, interests, and geographic location. Due to privacy issues, public benchmarks often lack user information. Some studies have explored how LLMs can infer relevant user profile information from browsing history and reviews (Liu et al., 2024; Richardson et al., 2023). A detailed user model and data can equip LLM to provide personalized solutions to the user in various domains, such as recommendation systems (Liu et al., 2024), prediction tasks (Li and Zhao, 2021), dialog systems, suspiciousness detection (Yang and Menczer, 2023), and personalized generation/classification (Salemi et al., 2023).

Both non-parametric (Salemi et al., 2023; Yang et al., 2023; Salemi et al., 2024; Richardson et al., 2023) and parametric approaches (Tan et al., 2024; Alaluf et al., 2024) have been employed to incorporate user model and data into the decision-making process of LLMs. However, when it comes to LLM agents, to our knowledge, there is no prior art on the personalized LLM agent, and we are the first to introduce a personal LLM agent benchmark built on top of TravelPlanner (Xie et al., 2024). Our approach to extending the generic benchmark to a personal one and our development and evaluation framework apply to other tasks.

3 Personal LLM Agent: TravelPlanner+

We present the TravelPlanner+ to evaluate the effectiveness of LLM agents in generating personalized travel itineraries based on the user model and the reference information, referred to as sole planning. This setting eliminates the need for tool calls, as agents no longer need to gather information from scratch using tools. This provides an opportunity to evaluate agents' planning skills solely (Xie et al., 2024). This reference information comprises detailed and essential data provided directly to the agents, including restaurants, accommodations, and attractions in the specified cities in the query.

In this work, we develop distinct user models to benchmark the effectiveness of open- and closed-source LLM agents and evaluate agents' planning skills in generating personalized plans. The benchmarking process involves using queries from TravelPlanner's validation and test splits and four planning strategies to craft multi-day (three-, five-, or seven-day) itineraries tailored for each user. Additionally, we curate personal queries to evaluate the performance of LLM agents in crafting plans that align with specific personal profiles and concepts.

3.1 User Model Generation Pipeline

We leverage the GPT-4 (Achiam et al., 2023) based AI User Model Generator (GPTs, 2024), which combines custom instructions and domain knowledge to generate user models for our travel plan generation. Refer to Table 8 for the prompt-related information and Appendix A.1 for sample-generated user models. We employ a structured representation of user models to consistently and effectively capture travel-related information, including interests, favorite cuisines, activities, and personal user concepts. We encapsulate the user models concisely to deal with the limited context length of the LLM. Human reviewers assess these synthetic user profiles to validate their accuracy and realism. For each synthesized user profile, the human reviewers manually verify that it aligns with expectations and closely mimics real-world human users. This process involves checking that each profile contains values for key user characteristics necessary to describe a user, including demographics, occupation, industry, and personal interests. We ensure that all fields are filled in by either filling in missing details based on realistic assumptions or removing them to save prompt tokens, ensuring efficiency. Additionally, we refine the generated personas to represent a balanced distribution of age groups, purchasing power, and ethnicity to reflect real-world diversity. The profiles are carefully curated to include various occupations and hobbies, ensuring the generated plans are personalized and varied across user profiles. Furthermore, we align user preferences with constraints specified in reference information, pushing the LLM's personalized plan creation capabilities. By smartly choosing preferences for user interests, we can test the model's ability to handle diverse and complex planning scenarios effectively. Additionally, we manually assign pet names to users who have pets. Appendix A.2 presents a comprehensive analysis of the user models.

Furthermore, we generated personal queries that contain user-model guided customized language. Figure 1 presents some examples of such customization: *pet-friendly* $\rightarrow$ *Goldie* | *Muffin*. Using this approach, we created 5 cuisine-based and 60 pet name-based personalized queries using simple replacements.

3.2 Personal LLM Agent

To integrate the user model into the LLM agent's decision-making process and to create a personal LLM agent, we choose a non-parametric approach due to its wide applicability and seamless integration into various use cases. We integrate a structured user model into four planning strategies: Direct, CoT (Wei et al., 2022), ReAct (Yao et al., 2023b), and Reflexion (Shinn et al., 2023). Along with injecting user models, we add key phrases into the prompt that guide the LLM in generating personalized plans that align with the target user model attributes. In Direct planning, the personal LLM agent creates personalized plans based on the system prompt, a one-shot example, the user model, and reference information.

CoT and ReAct strategies extend direct planning by encouraging step-by-step reasoning. CoT focuses on breaking down the problem into smaller steps, while ReAct incorporates detailed Thought and Action phases. Lastly, the Reflexion strategy employs a feedback loop and a scratchpad, enabling the LLM to evaluate and improve the plan iteratively. The specific user model contained prompts employed in these experimental settings are detailed in Appendix A.3 - Table 16 for reference. As a baseline, the generic LLM agent follows similar planning strategies without the user model information.

3.3 Prompt Improvements

Prompts are constructed to include strategy-specific wording, user queries, in-context example, and reference information in personal and generic settings, thereby providing meaningful context for generating effective plans. Compared to TravelPlanner's benchmark implementation (Xie et al., 2024), in addition to user model integration for personal LLM agent, we made several enhancements to the prompts (Refer to Appendix A.3). These enhancements involve:

- Adjusting in-context examples to exclude specific restaurant and accommodation names to avoid biasing the models
- The restaurants and accommodations from the TravelPlanner (Xie et al., 2024) benchmark are randomly assigned to various cities. Therefore, we replaced the names of restaurants and accommodations with anonymized names to mitigate any bias introduced by random assignment, as shown in the example in Table 10
- Substituting negative information with neutral or positive details to promote a positive outlook, enhancing model reasoning. For instance, we replaced accommodation rules such as 'No smoking' with 'Allows children under 10, allows parties, allows pets, and permits visitors.'

These enhancements significantly improved the generation of more effective plans compared to their pre-modification state; refer to Appendix A.8.

3.4 LLM-as-a-Judge

In this study, we employed LLM-as-a-Judge as an evaluation framework to serve as subjective tests, which have been demonstrated and proven effective in approximating human preferences (Chiang and Lee, 2023; Thomas et al., 2024; Chan et al., 2024; Zheng et al., 2023). We provided the LLM judge with the user model, which encapsulates user-related information, preferences, and personal concepts, along with a pair of generic and personal plans. The generic plans were generated by an LLM agent without access to the user model. In contrast, the personal plans were created by a personal LLM agent tailored to the user's specific needs. The LLM judge evaluated which plan matched the user's preferences and requirements. This test aims to measure the effectiveness of LLMs in creating highly tailored travel experiences that enhance user satisfaction. Our findings demonstrate that personalized travel plans significantly outperform generic ones in relevance and suitability, thereby validating the potential of personal LLM agents to deliver superior, user-centric travel solutions. More details are shown in Appendix A.6.

4 Experiments Setup and Results

4.1 LLM Agents

For this study, we evaluated both open-source and closed-source LLMs using various prompting strategies, including GPT-3.5-Turbo-16k (OpenAI, 2022) and Llama-3-8B-instruct-8k (AI@Meta, 2024). Each model selected could handle a large context window suitable for various planning strategies used in our experiments. For the open-source, we selected models in the $\leq$10B range to explore their suitability for customers' edge devices for designing a personal LLM agent on the edge.

4.2 Evaluation Strategies

In this study, travel plans are generated for the validation and test splits of TravelPlanner benchmark exploring two distinct settings:

- **Generic Setting**: In this setting, the LLM generates generic travel plans based solely on the query and necessary reference data. These plans are designed to adhere strictly to com-

Table 1: Performance indicators (%) of different LLM agents and planning strategies on the TravelPlanner validation set. The Personal plans are averaged over 5 user models. The best outcomes are in bold, and the second-bests are underlined. (Refer to Appendix A.8 for additional baseline numbers.)

Planning strategy	Generic plans						Personal plans					
	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro			Micro	Macro	Micro	Macro	
GPT-3.5-Turbo												
Direct	**100**	**67.15**	**3.33**	20.24	5.00	0	**100**$_{\pm 0.00}$	**65.67**$_{\pm 0.51}$	3.67$_{\pm 1.15}$	**24.12**$_{\pm 1.65}$	6.33$_{\pm 3.74}$	0.34$_{\pm 0.31}$
CoT	**100**	66.94	**3.33**	**20.95**	**9.44**	**1.11**	99.66$_{\pm 0.31}$	65.21$_{\pm 1.19}$	**5.00**$_{\pm 1.88}$	20.33$_{\pm 2.54}$	**6.67**$_{\pm 1.71}$	**0.59**$_{\pm 0.39}$
ReAct	**100**	64.44	2.22	9.28	2.78	0	99.89$_{\pm 0.25}$	59.00$_{\pm 0.92}$	1.56$_{\pm 1.14}$	4.62$_{\pm 1.90}$	1.00$_{\pm 0.82}$	0$_{\pm 0.00}$
Reflexion	**100**	63.47	0.56	3.57	1.11	0	99.44$_{\pm 0.00}$	64.14$_{\pm 1.06}$	1.34$_{\pm 0.50}$	9.76$_{\pm 1.52}$	2.67$_{\pm 1.64}$	0$_{\pm 0.00}$
Llama-3-8B-instruct												
Direct	**100**	**76.53**	**16.11**	**31.67**	**8.33**	1.67	**98.89**$_{\pm 0.00}$	**73.16**$_{\pm 1.03}$	**11.78**$_{\pm 1.73}$	**18.33**$_{\pm 2.69}$	**7.33**$_{\pm 2.68}$	**1.22**$_{\pm 0.72}$
CoT	98.89	69.65	8.33	16.43	5.00	**2.22**	**98.89**$_{\pm 0.00}$	68.32$_{\pm 0.95}$	4.67$_{\pm 1.87}$	11.71$_{\pm 2.02}$	4.78$_{\pm 1.09}$	0.89$_{\pm 0.50}$
ReAct	45.00	32.01	2.78	4.28	1.67	0	35.89$_{\pm 2.71}$	25.59$_{\pm 2.12}$	1.22$_{\pm 0.82}$	4.86$_{\pm 1.32}$	1.22$_{\pm 0.73}$	0.11$_{\pm 0.25}$
Reflexion	52.22	37.15	2.22	9.76	1.67	0.56	34.67$_{\pm 6.45}$	24.33$_{\pm 4.64}$	1.00$_{\pm 0.91}$	6.33$_{\pm 1.82}$	2.78$_{\pm 1.76}$	0.22$_{\pm 0.31}$

Table 2: Performance indicators (%) of LLama-3-8B-instruct LLM agent using Direct planning strategy on the TravelPlanner validation split for 20 user models.

Plans	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro	
Validation Split - Direct - 20 User Models						
Generic	100	76.53	16.11	31.67	8.33	1.67
Personal	98.89$_{\pm 0.00}$	72.90$_{\pm 1.14}$	10.56$_{\pm 2.27}$	18.87$_{\pm 2.53}$	7.50$_{\pm 2.91}$	1.56$_{\pm 0.84}$

Table 3: Performance indicators (%) of LLama-3-8B-instruct LLM agent using Direct and CoT on TravelPlanner test split for one randomly selected user model.

Plans	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro	
Direct - Test Split						
Generic	98.90	74.30	10.90	30.35	11.80	1.40
Personal	98.50	72.60	9.90	17.90	6.80	1.40
CoT - Test Split						
Generic	98.80	69.41	3.70	14.94	7.10	0.50
Personal	96.80	67.21	2.80	10.09	3.50	0.60

monsense and hard constraints specified in the queries without incorporating any user model. The metrics used are delivery rate, commonsense constraint pass rate, hard constraint pass rate, and final pass rate. For more information, refer to Appendix A.7.

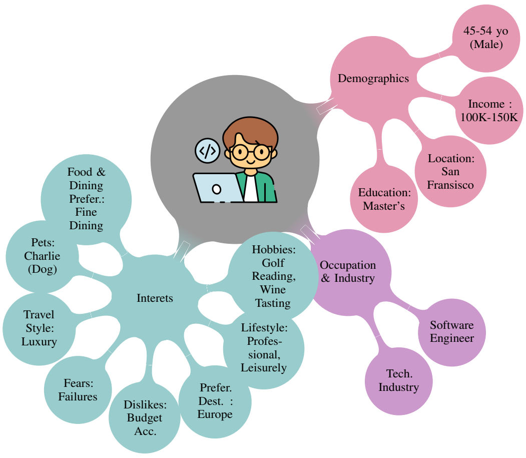
- **Personal Setting**: The plans are generated by incorporating the user model into the prompt and the details used in generic plan generation. This approach enables the LLM agent to create personalized plans from the LLM agent tailored to user-specific preferences and needs. To evaluate the impact of personalization on the generated plans, we employed the following metrics in addition to standard performance metrics:
  - **Preference Rate**: Measures the proportion of personal plans preferred over non-personal plans in percentage
  - **Reasoning Analysis**: Human evaluators examined the reasoning section provided by the LLM judge to understand why the LLM preferred one plan over the other

4.3 Experimental Results

In this section, we explore the personalization capabilities of various LLM agents. The results and observations are detailed as follows:

4.3.1 Performance Indicator Comparison

Table 1 details the performance of injecting user models into prompted LLM agents for generating personal plans compared to their generic (non-personal) counterparts. The performance metrics for personal plans are averaged across five randomly chosen user models to evaluate their efficacy on the validation split of the base dataset. Due to the computational requirement to run all the validation and test queries (1180*N in total, where N is the number of user models) being too heavy, we selected only five for validation purposes. Although the numbers are within a close range, the performance indicators for personal plans show a slight decline compared to generic plans for both the prominent models, GPT-3.5-Turbo and Llama-3-8B-instruct. For generic plans, the delivery rate of GPT-3.5-Turbo is 100% across all prompting strategies. However, due to its limited context length compared to GPT-3.5-Turbo, Llama-3-8B-instruct achieves 100% and 98.89% delivery rates only for the direct and CoT prompting strategies, respectively. When comparing constraint-based indicators for personal and generic plans, the Llama-3-8B-instruct model achieves the most optimal



Query ID	Reasoning
val-56	"The second plan has more varied food options, a longer stay in Orlando with multiple attractions, and no flight which might be stressful for the traveler."
val-72	"Luxury Traveler prefers fine dining and high-end accommodations, both plans provide suitable options, but Plan P has a more consistent high-end dining experience"
val-81	"The P plan has more luxurious accommodations, which suits the traveller's taste."
val-122	"Pet-friendly accommodation"
val-142	"Plan P accommodates traveller's preferences for European cities, fine dining, and luxury travel"

Figure 2: **Left**: User-related information, preferences, and personal concepts from User Model no. 2, **Right**: Reasoning examples from the LLM-as-a-Judge (Llama-3-8B-Instruct) explaining its preference for personal plans over generic plans for selected validation queries. Both types of plans were generated by the Llama-3-8B-Instruct agent. Icon by Freepik (Iconfromus).

commonsense, hard constraint, and final pass rates, making it a good option for on-device, locally deployed personal LLM agents. We also tested the Qwen-1.5-7B-chat model, but its results did not match the performance of Llama-3-8B-instruct, Refer to the Appendix A.9 for more information.

Direct and CoT prompting strategies demonstrated superior performance across all evaluated indicators compared to strategies that perform iterative refinements, such as ReAct and Reflexion. Delving into the failure cases of these strategies demonstrates their significant drawbacks in plan formation due to iterative action exploration within limited context windows under both generic and personal planning settings. Refer to Figure 3. Our empirical results align with (Xie et al., 2024; Verma et al., 2024; Hao et al., 2024) and suggest that in complex multi-constraint and long-term travel planning setting, the LLM agent faces challenges coordinating their actions with their analytical thinking in the ReAct and Reflexion strategies. Additionally, we observed that hallucinations in responses when extracting concrete plans from raw outputs of LLMs significantly hindered the effectiveness of ReAct and Reflexion. More information is provided in the Appendix A.5.

In light of the aforementioned observations, personal plans were generated for all 20 user models using the Llama-3-8B-instruct LLM agent with the direct prompting strategy. The results are highlighted in Table 2. Additionally, an evaluation was conducted on the test split of the base dataset for a randomly chosen user model for the two best-performing prompting strategies. The experimental results, shown in Table 3, are inline with the previously mentioned observations.

Table 4: Preference rates for plans generated by GPT-3.5-Turbo and Llama-3-8B-Instruct using various methods for validation split on 5 user models, with Llama-3-8B-Instruct and Gemma2-9B-Instruct as judges.

Method	Planner	GPT-3-Turbo		Llama-3-8B-Instruct	
	Judge	Llama3	Gemma2	Llama3	Gemma2
Direct	Generic	39.22	27.56	40.22	32.33
	Personal	**60.78**	**72.44**	**59.78**	**67.67**
CoT	Generic	40.23	33.1	37.44	32.67
	Personal	**59.77**	**66.9**	**62.56**	**67.22**
ReAct	Generic	44.00	42.56	43.81	27.14
	Personal	**56.00**	**57.44**	**56.19**	**72.86**
Reflexion	Generic	40.09	38.42	38.33	27.75
	Personal	**59.91**	**61.58**	**61.67**	**72.25**

4.3.2 Personal Setting Evaluations

The evaluation results on the preference rate in personal settings are shown in Table 4, including two judges, Llama-3-8B-instruct and Gemma2-9B-instruct. As shown, for both judges, personal travel plans were consistently preferred over non-personalized ones, with the preference for personal plans ranging from 56% to 72.86%. While extending to 20 personas, the Direct method reaches 61.06% and 74.4%. For the results generated on the test set by Llama-3-8B-instruct, the preference rates from the two judges reach 66.5% and 87% for Direct, 72.2% and 87.3% for CoT, respectively. This demonstrates that when the personal LLM agent tailors travel plans to specific users, the relevance and suitability of these plans significantly increase, aligning more closely with individual preferences and needs. The LLM judge also provides detailed reasoning for its selections, highlighting the factors contributing to its decisions. We converted the results into word cloud in Appendix A.6.

Figure 2 shows a sample user model, highlighting the diverse range of user-related information.The LLM judge selects its preferred personalized or generic plan and justifies its choice by considering the user model. The reasoning provided by the LLM judge is displayed for various queries, illustrating the detailed attention it pays to the alignment of travel plans with the user model. For instance, the judge emphasizes luxury accommodations, fine dining experiences, and pet-friendly options, all tailored to the user's preferences.

4.3.3 Personal Queries

We constructed several personal queries based on the approach described in Section 3.1 to evaluate the preference rate and reasoning. The results demonstrate that the preference rate for personal plans reaches 58.5% and 86.2%, respectively, according to the evaluations by the two judges. The significant preference rates achieved underscore this personalization strategy's effectiveness in enhancing travel plans' relevance and satisfaction. For more information, refer to the Appendix A.4.

5 Conclusions

Our work introduced TravelPlanner+, a personalized version of the TravelPlanner benchmark. We created user models to integrate into the decision-making process of the LLM agent. This was the first study entirely devoted to personal LLM agents. Our design decisions, such as the size of the open-source models and the non-parametric approach to personalization, were made to facilitate the on-device deployment of LLM agents, providing a privacy-preserving solution by design. As demonstrated by the LLM-as-a-judge, our evaluation framework clarified the quality of individual plans, which were previously obscured by generic performance indicators.

6 Limitations

Our solution has some limitations as our first attempt to build a personal LLM agent benchmark. One limitation is the distribution gap between the synthesized user models and the actual data. Moreover, the size of the user models needs to be larger to capture the population unbiasedly. However, we need to mention that the goal of a personal LLM agent is to build a biased LLM agent towards the user. Still, by increasing the population size, we can further investigate the adverse effect of bias in the LLM agents and offer solutions to mitigate that.

Moreover, we used LLM-as-a-Judge to compare the generic and personal plans. Human evaluation is still necessary to fully assess the quality and alignment of personal plans with the user models. It would be interesting to explore the correlation between human assessments and the LLM-as-a-Judge for this specific application to further support the validity of this approach, as indicated by other research. Examining the differences between various LLM judges is an intriguing research direction that we aim to pursue in future iterations of this work.

Additionally, we did not investigate quantized models in this study. The choice of task is also limited to travel planning, but we anticipate the findings to be translatable to other tasks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2024. Myvlm: Personalizing vlms for user-specific queries. *arXiv preprint arXiv:2403.14599*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Featured GPTs. 2024. Ai user persona generator. Accessed: 2024-07-08.

Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. 2024. Large language models can plan your travels rigorously with formal verification tools. *arXiv preprint arXiv:2404.11891*.

Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*.

Sheng Li and Handong Zhao. 2021. A survey on representation learning for user modeling. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4997–5003.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Dianshu Liao, Shidong Pan, Qing Huang, Xiaoxue Ren, Zhenchang Xing, Huan Jin, and Qinying Li. 2023. Context-aware code generation framework for code repositories: Local, global, and third-party library awareness. *arXiv e-prints*, pages arXiv–2312.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 452–461.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. *arXiv preprint arXiv:2404.05970*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. On the brittle foundations of react prompting for agentic large language models. *arXiv preprint arXiv:2405.13966*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and

Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an ai-powered malicious social botnet. *arXiv preprint arXiv:2307.16336*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *arXiv preprint arXiv:2401.07339*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 User Model Sample

User models capture a high-level overview of the user. We used a structured format to capture the travel-related information. The examples of generated user models 2 and 9 are shown in Table 5 and Table 6, respectively.

Table 5: Sample of generated user model 2

```
Demographics:
- Age Range: 45-54
- Gender: Male
- Income Level: $100,000-$150,000
- Location: San Francisco, USA
- Education: Master's Degree

Occupation & Industry:
- Job Title: Software Engineer
- Industry Type: Technology

Interests
- Hobbies: Golf, Reading, Wine
Tasting
- Lifestyle: Professional, Leisurely
- Preferred Destinations: European
Cities
- Food and Dining Preferences: Fine
Dining
- Dislikes: Budget accommodations
- Fears: Failure
- Pets: Charlie, Dog
- Travel Style: Luxury Traveler
```

Table 6: Sample of generated user model 9

```
Demographics:
- Age Range: 18-24
- Gender: Female
- Income Level: <$20,000
- Location: Buenos Aires, Argentina
- Education: High School Diploma

Occupation & Industry:
- Job Title: Barista
- Industry Type: Hospitality

Interests:
- Hobbies: Dancing, Social Media,
Traveling
- Lifestyle: Fun-loving, Budget-
conscious
- Preferred Destinations: Beach
Resorts
- Food and Dining Preferences: Street
 Food
- Dislikes: Boredom
- Fears: Missing out (FOMO)
- Pets: Luna, Cat
```

A.2 User Model Analysis

In this section, we analyze the distribution of different attributes of the 20 user models we have generated using the process outlined in Section 3.1. The overall schema of each user model is depicted in Figure 2. The prompt for user model generation is shown in Table 8. Table 7 presents each of the user model's categories, their associated sub-categories, and a count of all the unique sub-category values. Please note that the data is biased due to the size of the personas. We can increase the sample size and incorporate data from diverse sources to mitigate the bias from the limited number of personas. Regularly refine personas with real-world data and expert reviews to ensure balanced representation.

Demographics: Under Demographics, we notice that most of the users are in the age range of 25-44 (8), have an annual income of $70,000 - $120,000 (6), and hail from North America (7).

Occupation & Industry: Another facet that heavily influences travelling is the occupation of the users. For instance, unlike entrepreneurs, teachers or students would have specific holiday seasons. For this reason, we created highly diverse occupations with some industry overlap to accommodate the evaluation of generalizable travel agents. An agent who obtains good plans for all 19 occupations would be robust toward occupation diversity, which is the reality of current times.

Interests: This category encompasses the majority of personalization attributes. We cover hobbies, lifestyles, travel styles, pets, destinations, and dining preferences, as well as dislikes and fears. We consider that users can have multiple hobbies, but most (12) have travelling as their hobby. The rest of their interests are highly diverse, ranging from solo activities (e.g., reading) to group activities (e.g., music festivals) and from indoor activities (e.g., gaming) to outdoor activities (e.g., hiking). Most of the other interests attributes follow similar trend of occupation and cover a multitude of options. The pets aspect is limited to a small number of people because we extrapolate that the difficulties involved in travelling with pets would often discourage people from doing so.

Table 7: Count-based analysis of profile categories and their sub-categories for attributes of all the 20 personas.

User Model Categories	Sub-categories	Sub-categories values for ($P = 20$) Personas
Demographics	Age Range	25-34 (4), 35-44 (4), 45-54 (3), 18-24 (3), 55-64 (2), 50-59 (2), 40-49 (1), 30-39 (1)
	Gender	Female (10), Male (9), Non-binary (1)
	Income Level	$90,000-$120,000 (3), $70,000-$90,000 (3), $60,000-$80,000 (2), $40,000-$60,000 (2), $80,000-$100,000 (2), <$20,000 (2), $50,000-$70,000 (1), $120,000-$150,000 (1), $100,000-$120,000 (1), $100,000-$150,000 (1), <$30,000 (1), $30,000-$50,000 (1)
	Location	North America (7), Asia (5), Europe (5), Australia (2), South America (1)
	Education	Bachelor's Degree (11), Master's Degree (5), High School Diploma (2), PhD (1), Currently in University (1)
Occupation & Industry	Job Title	Software Developer (2), Marketing Manager (1), Product Manager (1), Art Curator (1), Retired Teacher (1), Film Producer (1), Financial Analyst (1), Part-time Retail Worker (1), UX Designer (1), Entrepreneur (1), Digital Nomad (1), Software Engineer (1), HR Manager (1), Student (1), Graphic Designer (1), Retired (1), Journalist (1), Real Estate Agent (1), Barista (1)
	Industry Type	Technology (3), Education (2), Media (2), Advertising (1), Electronics (1), Museum (1), Entertainment (1), Finance (1), Retail (1), Manufacturing (1), Freelance (1), Corporate (1), Not Available (1), Real Estate (1), IT (1), Hospitality (1)
Interests	Hobbies	Traveling (12), Reading (5), Yoga (3), Wine Tasting (3), Photography (2), Gardening (2), Gaming (2), Cooking (2), Fishing (2), Hiking (1), Video Games (1), Anime (1), Art Collecting (1), Live Music (1), Biking (1), Coding (1), Pilates (1), Fine Dining (1), Skiing (1), K-pop (1), Design (1), Business Networking (1), Blogging (1), Golf (1), Music Festivals (1), Art (1), Volunteering (1), Painting (1), Surfing (1), Writing (1), Museum Visits (1), Language Learning (1), Boating (1), Movies (1), Dancing (1), Social Media (1)
	Lifestyle	Social (4), Tech-savvy (3), Health-conscious (2), Intellectual (2), Relaxed (2), Professional (2), Creative (2), Family-oriented (2), Active (2), Structured (1), Innovative (1), Sophisticated (1), Community-oriented (1), Glamorous (1), Affluent (1), Trendy (1), Minimalist (1), Busy (1), Strategic (1), Independent (1), Flexible (1), Leisurely (1), Balanced (1), Eco-conscious (1), Budget-oriented (1), Curious (1), Fun-loving (1), Budget-conscious (1)
	Travel Style	Not Available (10), Cultural Explorer (3), Luxury Traveler (3), Adventure Seeker (1), Solo Traveler (1), Family Traveler (1), Backpacker (1)
	Preferred Destinations	Exotic Islands (2), National Parks (1), Tech Expos (1), Historic Cities (1), Music Festivals (1), Quiet Countryside (1), Luxury Resorts (1), Major Cities (1), Design Capitals (1), Business Hubs (1), Remote Locations (1), European Cities (1), Family-friendly Resorts (1), Not Available (1), Coastal Areas (1), Countryside (1), Historical Sites (1), Caribbean Islands (1), Tech Conferences (1), Beach Resorts (1)
	Food and Dining Preferences	Organic (2), Home-cooked (2), Farm-to-table (2), Seafood (2), Vegan (1), Sushi (1), Ramen (1), French Cuisine (1), BBQ (1), Craft Beer (1), Healthy (1), Gourmet (1), Fast Food (1), Scandinavian Cuisine (1), Traditional Indian (1), Vegetarian (1), Fine Dining (1), Asian Cuisine (1), Not Available (1), Local Cuisine (1), Latin Cuisine (1), Spicy Food (1), Street Food (1)
	Dislikes	Fast food (2), Budget accommodations (2), Pollution (1), Crowded places (1), Bureaucracy (1), Mass tourism (1), Long work hours (1), Crowds (1), Studying (1), Clutter (1), Inefficiency (1), Restrictions (1), Unpredictability (1), Wastefulness (1), Conformity (1), City noise (1), Cold Weather (1), Long commutes (1), Boredom (1)
	Fears	Heights (1), Stagnation (1), Losing cultural heritage (1), Career stagnation (1), Health problems (1), Public Failure (1), Economic instability (1), Job insecurity (1), Creative block (1), Business failure (1), Isolation (1), Failure (1), Job loss (1), Climate Change (1), Monotony (1), Health issues (1), Ignorance (1), Economic Downturn (1), Job instability (1), Missing out (FOMO) (1)
	Pets	Not Available (14), Dog (4), Cat (2)

A.3 Prompts

Since we adopted a non-parametric approach, i.e. prompt engineering, for using LLMs as travel agents, we have dedicated this section to providing access to our prompts, which support the reproducibility of our work. We recommend reviewing the prompts starting from Table 16.

System Start Prompt: In accordance with the popular prompt-based approaches, we begin using a System Start prompt. The starting prompt used for Direct and CoT is outlined in Table 9. The thought prompt outlined in Table 14 under the Thought prompt has been curated for particular usage during the thinking phase of the ReAct and Reflexion-based agents. Overall, this prompt segment outlines the agent's scope and purpose and some of the rules for response generation.

Special Instruction: For CoT (13), ReAct (14), and Reflexion (14), we provide additional instructions before the one-shot example segment to further match the agent behavior to the planning strategy of each system respectively.

One Shot Example: We provided a one-shot example for all the systems to guide the agents towards the expected style of generated travel plan. To reduce copy-paste mistakes based on the one-shot example, we anonymize the information present in the example using Xs (e.g., restaurant_XXXX). One-shot example prompt used for Direct, CoT, and ReAct can be found under Table 10. For Reflexion, we use different examples for the thought and action phases as highlighted in Table 11 and Table 12. Due to the context length limitations, we restricted our experiments to only a one-shot setting.

Historical Context: The differentiation factor of Reflexion is its ability to reflect on feedback about past turns. To this extent, an example of injection of past experiences into the prompt is presented in Table 15.

Core Query Information: We provide the outcome of Oracle tool usage (Given Information/Reference Information), chosen user model (user model), and the query (original or personalized) to all the agents. The ReAct and Reflexion systems receive additional context in scratch pads containing the past ten thought and action outcomes. The key difference between our personalized and non-personalized systems is the inclusion of the user model segment. Furthermore, we replace the query with a personalized one and evaluate the performance of both personalized (with user model) and generic (without user model) systems on query-guided personalization.

System End Prompt: Finally, we end the prompts for each system differently. The simplest is the Direct setting, which has no capability for additional reasoning. For the CoT, we append the traditional "Let's think step by step" instruction into the prompt. Although ReAct and Reflexion share a similar ending, which requires the agent to reason using the explicitly thought phase and choose an action from pre-defined actions (Table 14), we further include the past reflections for the Reflexion prompt.

Combining the prompt segments from each of the different tables in the order mentioned in Table 16, we create the final prompt for all the various planning strategies. We use the same prompt for all of our language models to ensure fairness in evaluation.

A.4 Generated Plans

As explained before we tested our approach for both generic and personal queries. The following subsections closely examine the generated plans in different settings.

A.4.1 Generic Queries: Generic vs. Personal Plans

Table 17 displays the generic plan generated for the generic query, while Table 18 presents the personal plan tailored to the same query for user model 2. The personal plan aligns more closely with the user's preferences for Italian and French cuisines and avoids repeating restaurants, ensuring a varied dining experience. It also provides detailed cost information, which helps manage the budget effectively. Additionally, the personal plan ensures all constraints are met, including valid restaurants and attractions, while the generic plan repeats a restaurant, indicating less thoughtful planning. The accommodations in the personalized plan better suit the user's luxury travel style, enhancing overall satisfaction.

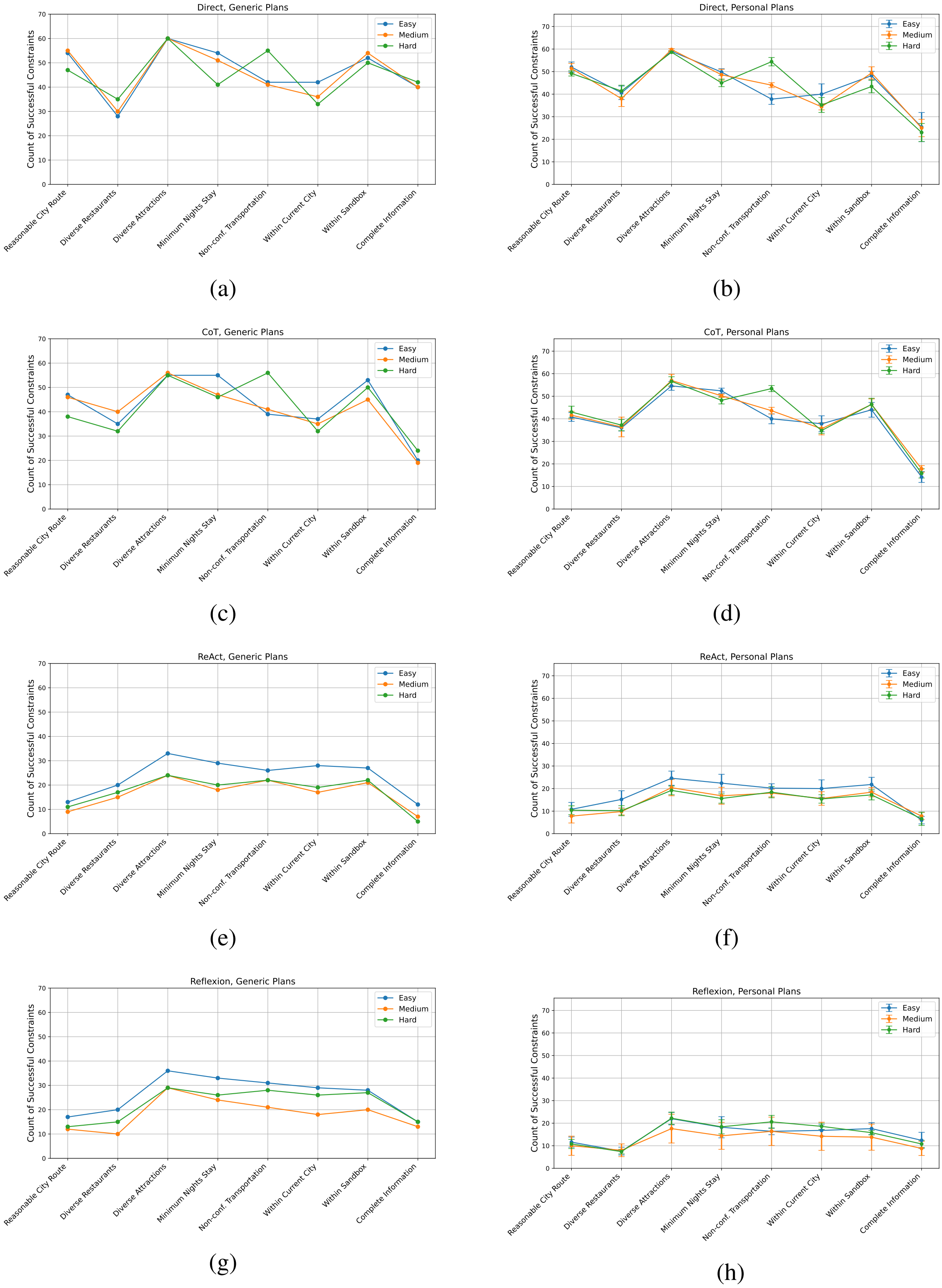


Figure 3: This figure illustrates the performance of various planning strategies, Direct, CoT, ReAct, and Reflexion, in fulfilling travel planner constraints across different difficulty levels on a validation set using Llama-3-8B. The left column represents the generic LLM agent plans, while the right column shows the metric values of the personal LLM agent for five user models. The count of successful constraints is plotted for each category, demonstrating the effectiveness of each approach under varying difficulty conditions. Error bars indicate standard deviations on the personal plans.

Table 8: User model generation prompt

```
You are an expert. Complete N distinct and diverse user personas in a structured
format requested as follow. Choose specific answers to each of the fields. For the
fields that are optional, you can randomly choose to fill them or remove them from
the persona. Remove the optional flag from the field.

Persona:

   Demographics:

      Age Range:

      Gender:

      Income Level:

      Location:

      Education:

   Occupation & Industry:

      Job Title:

      Industry Type:

   Interests:

      Hobbies:

      Lifestyle:

      Preferred Destinations [optional]:

      Food and Dining Preferences [optional]:

      Dislikes:

      Fears:

      Pets [optional]:

      Travel Style [optional]:

Think critically step by step to create a user persona.
```

A.4.2 Personal Queries: Generic vs. Personal Plans

Table 19 provides the generic plan example for the personal query, whereas Table 20 illustrates the personal plan example for the personal query for the user model 9. The user model 9 is shown in Table 6. The personal plan is better than the generic one since it prioritizes pet-friendly accommodations, recognizing Luna, user model 9's pet cat. Additionally, the personal plan includes more detailed and specific attractions, accommodations, and meals, ensuring a more enjoyable and comprehensive experience. In contrast, the generic plan overlooks the need for pet-friendly options and fails to select pet-friendly accommodations.

A.5 Issues in response with ReAct and Reflexion prompting

When generating plans using the ReAct and Reflexion strategies, which are more complex than the straightforward Direct and CoT planning strategies as seen in Figure 3, we encountered several issues:

- **Iterative Activity in Action Exploration:** We expected that the iterative refinement of ReAct and Reflexion strategies would help with the planning. However, our observations do not align with our expectations. Planning each day's itinerary with ReAct and Reflexion strategies, constrained by the limited context window of the models, makes it challenging

Table 9: System prompt that is used at the beginning of the prompt for Direct and CoT systems

```
You are a proficient planner with a keen understanding of personal preferences and
styles. Based on the provided information, persona, and query, please give me a
detailed and personalized plan, including specifics such as flight numbers (e.g.,
F0123456), restaurant names, and accommodation names. Note that all the information
in your plan should be derived from the provided data and aligned with the persona
details. You must adhere to the format given in the example. Additionally, all
details should align with common sense. The symbol `-' indicates that information is
 unnecessary. For example, in the provided sample, you do not need to plan after
returning to the departure city. When you travel to two cities in one day, you
should note it in the 'Current City' section as in the example (i.e., from A to B).
Always prioritize the query constraints first, especially when they conflict with
personal preferences. Incorporate personal preferences as secondary considerations.
```

to produce complete and coherent full-length plans. This issue is particularly pronounced for longer itineraries, such as those spanning five to seven days. Refer to Figure 3-(e to h), where the counts of the successful constraints are lower in all categories compared to Direct and CoT approaches (Figure 3-(a to d)). Our empirical results align with (Xie et al., 2024)in which they report agents' struggle to synchronize their actions with their analytical reasoning in the Reflexion strategy. More recent studies, (Verma et al., 2024), question the true capabilities of iterative refinement strategies such as ReAct. They suggest that ReAct's performance is not due to "interleaving reasoning trace with action execution". Instead, LLM's performance in sequential decision-making tasks like travel planning is due to the high similarity between exemplar problems and the query task. (Kambhampati et al., 2024) questions LLM's planning capabilities and suggests that LLMs can play a more vital role in a Generate-Test-Critique loop, with the LLM generating candidate plans and a bank of critics critiquing the candidate. Human as a critique in a loop has been applied to the Travelplanner successfully (Hao et al., 2024). These suggest that simpler strategies such as Direct and CoT are more suitable for complex multi-constraint and long-horizon travel planning tasks.

- **Hallucinations:** Various forms of hallucination were observed in the LLM responses. These included generating content beyond the provided reference information, producing plans for all seven days when only a single day's plan was requested at each step, and failing to adhere to the required output structure. Due to the generative nature of the models, they often failed to produce outputs that matched the exact patterns required for successful regex-based extraction.

The plots in Figure 3 illustrate the constraint adherence capabilities of different prompting approaches. Notably, both ReAct and Reflexion exhibit a marked drop in performance concerning the 'within the sandbox environment' constraint, reflecting instances of hallucination. Additionally, there are other failure cases not captured by these plots. By closely examining the LLM responses, we can identify these instances. To illustrate these issues further, we provide an example.

Table 21 illustrates a scenario where both issues mentioned above were encountered while using the ReAct prompting strategy with the Llama-3-8B-instruct LLM agent. In ReAct, each step involves Thought and Action sub-steps. While the Thought prompt specifically instructs the LLM to generate plans for only a single day at a time, the LLM erroneously generated thoughts for each day of the entire plan. Additionally, in the Action prompt, where a particular output structure is required (e.g., **CostEnquiry [Sub Plan]** and **Finish [Final Plan]**), the LLM struggled to consistently match this structure. In this example, the response to the Thought prompt generated plans for all seven days, causing the subsequent Action prompt to become too large to fit within the context window of the LLM agent.

A.6 LLM-as-a-Judge

A.6.1 Preference evaluation prompt

The prompt example in the preference evaluation is shown in Table 22. This evaluation method allows for a systematic comparison between generic and

Table 10: Anonymized one-shot example used for guiding the model to generate responses in the style of example plan.

```
***** Example *****

Query: Could you create a travel plan for 7 people from Ithaca to Charlotte spanning
 3 days, from March 8th to March 14th, 2022, with a budget of $30,200?

Personalized Travel Plan:
Day 1:
Current City: from Ithaca to Charlotte
Transportation: Flight Number: F3633413, from Ithaca to Charlotte, Departure Time:
05:38, Arrival Time: 07:46
Breakfast: restaurants_XXXX, Charlotte
Attraction: The Charlotte Museum of History, Charlotte
Lunch: restaurants_XXXX, Charlotte
Dinner: restaurants_XXXX, Charlotte
Accommodation: accommodations_XXXX, Charlotte

Day 2:
Current City: Charlotte
Transportation: -
Breakfast: restaurants_XXXX, Charlotte
Attraction: The Mint Museum, Charlotte; Romare Bearden Park, Charlotte
Lunch: restaurants_XXXX, Charlotte
Dinner: restaurants_XXX, Charlotte
Accommodation: accommodations_XXX, Charlotte

Day 3:
Current City: from Charlotte to Ithaca
Transportation: Flight Number: F3786167, from Charlotte to Ithaca, Departure Time:
21:42, Arrival Time: 23:26
Breakfast: restaurants_XX, Charlotte
Attraction: Books Monument, Charlotte
Lunch: restaurants_XXXX, Charlotte
Dinner: restaurants_XXXX, Charlotte
Accommodation: -

***** Example Ends *****
```

personalized plans, highlighting the impact of user-specific data on the planning process.

A.6.2 Word Cloud of Reasoning

The word clouds from the judges Llama-3-8B-Instruct and Gemma2-9B-Instruct on the plans generated by Llama-3-8B-Instruct and GPT-3.5-Turbo are shown in Figures 4 to 9. User-specific terms like "align" and "traveler preference" stand out, along with more general terms like "food options" and "attraction."

A.6.3 Preference Rate Evaluation

The four methods' evaluation results on the preference rate are shown in Tables 10 to 13, including two judges, Llama-3-8B-instruct and Gemma2-9B-instruct. For both judges, personal travel plans were consistently preferred over non-personalized ones for all four methods.

A.7 Generic Performance Indicators

The generic setting performance indicator provides baseline values for evaluating the LLM's performance in planning multi-day itineraries, independent of the user models. We assess the planning quality of the LLM agent using the following metrics as proposed in the TravelPlanner (Xie et al., 2024):

- **Delivery Rate**: Evaluates if the agent can deliver a plan within 30 steps
- **Commonsense Constraint Pass Rate**: Measures if the agent incorporates commonsense (across eight dimensions) into the plans
- **Hard Constraint Pass Rate**: Checks if the agent meets the hard requirements specified in the query
- **Final Pass Rate**: The proportion of plans that satisfy all the above indicators

Table 11: One-shot example used for the Thought phase of ReAct and Reflexion systems to exemplify the nature of thoughts required for travel planning that abides by commonsense constraints

```
***** Example *****

Day 1: The first day involves traveling from Ithaca to Charlotte. Considering an
early morning flight will maximize the day in Charlotte. Breakfast can be planned
upon arrival, followed by visiting a popular attraction. Lunch and dinner should be
at well-reviewed local restaurants, and accommodation should be comfortable and
centrally located.

***** Example Ends *****
```

Table 12: One-shot example for ReAct and Reflexion system that guides their cost inquiry action at a single turn

```
***** Example for CostEnquiry action *****

{
  "people_number": 7,
  "day": 1,
  "current_city": "from Ithaca to Charlotte",
  "transportation": "Flight Number: F3633413, from Ithaca to Charlotte, Departure
  Time: 05:38, Arrival Time: 07:46",
  "breakfast": "restauarant_23, Charlotte",
  "attraction": "The Charlotte Museum of History, Charlotte",
  "lunch": "restaurants_814, Charlotte",
  "dinner": "restaurants_128, Charlotte",
  "accommodation": "accomodation_210, Charlotte"
}

***** Example Ends *****
```

A.8 Effect of Prompt modification

The improvements detailed in Section 3.3 to the base prompts used in TravelPlanner resulted in enhanced performance metrics, as illustrated in Table 23. We compared the performance of the original and improved prompts with Llama-3-8B-instruct using the Direct and CoT prompting strategies. Performance improved substantially across all metrics considered, especially the Commonsense Macro pass rate and Hard Constraint Micro pass rates, which more than doubled after these enhancements were implemented.

The comparison of sole-planning in the TravelPlanner validation split to TravelPlanner+ with modified prompts for GPT-3.5-Turbo is presented in Table 24. The metrics for the TravelPlanner prompts are sourced from the base paper and compared to the generic planning values of TravelPlanner+. Notably, the improved prompts achieved a 100% delivery rate across all four prompting strategies. Additionally, significant improvements were observed in nearly all other performance indicators with the enhanced prompting style.

A.9 Qwen-1.5-7B-chat model results

We also experimented with another open-source model, Qwen-1.5-7B-chat, which features a significantly larger context window of 128K tokens than the 8K tokens in the Llama-3-8B-instruct model. Table 25 mentions the results using this model. Qwen-1.5-7B-chat consistently achieves a 100% delivery rate across all prompting strategies for personalized plans. However, while this model successfully generates all plans due to its extensive context length, its adherence to constraints is comparatively less effective across nearly all strategies when compared to other models.

Table 13: CoT system is guided to display thorough analysis through step-by-step reasoning

```
Break down the instructions into a sequence of logical steps that build upon each
other to guide the planner through creating a personalized plan. Each step should
follow from the preceding one, leading the planner to consider all necessary details
 systematically, ensuring that all details are logically sequenced and align with
requested constraints.
```

Table 14: ReAct and Reflexion systems are requested to split their reasoning into explicit thought and action phases with different system prompts for each phase. Additionally, the Reflexion system can access the past 10 thought and action outcomes to provide additional natural language feedback.

```
--- Thought Prompt ---

You are a proficient planner with a keen understanding of queries and personal
preferences. Based on the provided query, information, and persona, please give me
concise thoughts on how to solve this task. Note that the information in your
thought should be derived from the provided data and aligned with the query
constraints and persona details. Additionally, all details should align with
commonsense constraints. Attraction visits and meals are expected to be diverse. The
 `Thought' should involve concise reasoning about the steps. Don't provide any
action in this step. Always prioritize the query constraints first, especially when
they conflict with personal preferences. Incorporate personal preferences as
secondary considerations.

--- Action Prompt ---

The `Action' phase should consist of planning, that can be only one of two types:

- CostEnquiry[Sub Plan]: This function is used to calculate the cost of a detailed
sub plan, which you need to input the people number and plan in JSON format. The sub
 plan should encompass a complete one-day plan. An example will be provided for
reference. Don't use null for the information that is unnecessary inside sub-plan,
use `-' string instead.
- Finish[Final Plan]: Use this function to indicate the completion of the task. You
must submit a final, complete plan as an argument.

At a time, only one function should be called. If CostEnquiry is called, it should
only contain the sub-plan for a single day. Ensure that CostEnquiry calls only
include the same keys as in the provided example. Do not add any extra keys beyond
those shown. New information might be added mid-planning based on earlier thoughts
and actions. Adjust the plan accordingly, but always ensure each action pertains to
a single day or calls Finish if the plan for all requested days is concluded.
```

Table 15: Our Reflexion adaptation uses this prompt to generate natural language feedback to understand and explain the errors made during plan generated so far.

```
You are an advanced reasoning agent that can improve based on self refection. You
will be given a previous reasoning trial in which you were given access to an
automatic cost calculation environment, a travel query to give plan, a user  and
relevant information. Only the selection whose name and city match the given
information will be calculated correctly. You were unsuccessful in creating a plan
because you used up your set number of reasoning steps. In a few sentences, Diagnose
 a possible reason for failure and devise a new, concise, high level  plan that aims
 to mitigate the same failure. Use complete sentences.

Given information: {text}

Previous trial:
Query: {query}{scratchpad}

Reflection:
```

Table 16: Summary of similarities and differences of prompt creation for each of the four systems. Combining the prompts from tables referenced in each column would generate a final prompt for that system. The text or instruction mentioned in the cells should be used in the cells without any table reference.

Prompt Segment (↓)	**Planning Strategy** (→)			
	Direct	**CoT**	**ReAct**	**Reflexion**
System Start	Table 9	Table 9	Table 14 - Thought Prompt or Table 9	Table 14 - Thought Prompt or Table 9
Special Instructions	✗	Table 13	Table 14	Table 14
One Shot Example	Table 10	Table 10	Thought - Table 11 Action - Table 12	Thought - Table 11 Action - Table 12
Given information:	Oracle tool output			
User Model	Formatted user model (Example in Table 5)			
Query	*query*	*query*	*query scratchpad*	*query scratchpad*
System End	Personalized Travel Plan:	Let's think step by step. Personalized Travel Plan:	You must use Finish to indicate you have finished the task. And each action only calls one function once. Now, based on your reasoning in the Thought phase, please proceed with only one necessary action from CostEnquiry or Finish.	You must use Finish to indicate you have finished the task. And each action only calls one function once. {*historical reflections*} Now, based on your reasoning in the Thought phase, please proceed with only one necessary action from CostEnquiry or Finish.

Table 17: Generic plan of sample validation query (val-74) by Llama-3-8B-Instruct with direct method for user model no. 2 (shown in Fig. 2)

```
Query: "Can you create a travel plan for a group of 5 departing from Charlotte
heading to Hilton Head, to be carried out over 3 days, from March 26th to March 28th
, 2022? The budget for this trip is capped at $7,000. We have a preference for
Italian and French cuisines during our trip."
```


```
[
    {
        "days": 1,
        "current_city": "from Charlotte to Hilton Head",
        "transportation": "Flight Number: F4055090, from Charlotte to Hilton Head,
        Departure Time: 18:07, Arrival Time: 19:27",
        "breakfast": "Name: Hunger's Hub, Cuisines: Tea, Bakery, Desserts, Seafood,
        Average Cost: 97, Aggregate Rating: 3.1, Hilton Head",
        "attraction": "Coastal Discovery Museum, Hilton Head",
        "lunch": "Name: Ashirbad, Cuisines: Pizza, Bakery, Mediterranean, Seafood,
        Average Cost: 50, Aggregate Rating: 3.6, Hilton Head",
        "dinner": "Name: Wrapster, Cuisines: French, Seafood, Desserts, Italian,
        Average Cost: 18, Aggregate Rating: 3.2, Hilton Head",
        "accommodation": "Name: Hip, Vibrant, COLORFUL Downtown Manhattan 1 Bed,
        room type: Entire home/apt, price: 763.0, minimum nights: 2.0, review rate
        number: 1.0, house_rules: No visitors, maximum occupancy: 7, Hilton Head"
    },
    {
        "days": 2,
        "current_city": "Hilton Head",
        "transportation": "-",
        "breakfast": "Name: Hunger's Hub, Cuisines: Tea, Bakery, Desserts, Seafood,
        Average Cost: 97, Aggregate Rating: 3.1, Hilton Head",
        "attraction": "Harbour Town Lighthouse, Hilton Head; Coligny Beach Park,
        Hilton Head",
        "lunch": "Name: Ashirbad, Cuisines: Pizza, Bakery, Mediterranean, Seafood,
        Average Cost: 50, Aggregate Rating: 3.6, Hilton Head",
        "dinner": "Name: Wrapster, Cuisines: French, Seafood, Desserts, Italian,
        Average Cost: 18, Aggregate Rating: 3.2, Hilton Head",
        "accommodation": "Name: Hip, Vibrant, COLORFUL Downtown Manhattan 1 Bed,
        room type: Entire home/apt, price: 763.0, minimum nights: 2.0, review rate
        number: 1.0, house_rules: No visitors, maximum occupancy: 7, Hilton Head"
    },
    {
        "days": 3,
        "current_city": "from Hilton Head to Charlotte",
        "transportation": "Flight Number: F4056985, from Hilton Head to Charlotte,
        Departure Time: 20:07, Arrival Time: 21:31",
        "breakfast": "Name: Hunger's Hub, Cuisines: Tea, Bakery, Desserts, Seafood,
        Average Cost: 97, Aggregate Rating: 3.1, Hilton Head",
        "attraction": "Books Monument, Charlotte",
        "lunch": "Name: Ashirbad, Cuisines: Pizza, Bakery, Mediterranean, Seafood,
        Average Cost: 50, Aggregate Rating: 3.6, Hilton Head",
        "dinner": "Name: Wrapster, Cuisines: French, Seafood, Desserts, Italian,
        Average Cost: 18, Aggregate Rating: 3.2, Hilton Head",
        "accommodation": "-"
    }
]
```

Table 18: Personal plan of val-74 by Llama-3-8B-Instruct with Direct method for User Model 2

```
Query: "Can you create a travel plan for a group of 5 departing from Charlotte
heading to Hilton Head, to be carried out over 3 days, from March 26th to March 28th
, 2022? The budget for this trip is capped at $7,000. We have a preference for
Italian and French cuisines during our trip."
```



```
[
    {
        "days": 1,
        "current_city": "from Charlotte to Hilton Head",
        "transportation": "Flight Number: F4055090, from Charlotte to Hilton Head,
        Departure Time: 18:07, Arrival Time: 19:27, Cost: 460",
        "breakfast": "-",
        "attraction": "-",
        "lunch": "-",
        "dinner": "Name: Wrapster, Cuisines: French, Seafood, Desserts, Italian,
        Average Cost: 18, Aggregate Rating: 3.2, Hilton Head",
        "accommodation": "Name: Rustic room in renovated Bushwick apartment, room
        type: Private room, price: 309.0, minimum nights: 1.0, review rate number:
        2.0, house_rules: No smoking, maximum occupancy: 1, Hilton Head"
    },
    {
        "days": 2,
        "current_city": "Hilton Head",
        "transportation": "-",
        "breakfast": "Name: Hunger's Hub, Cuisines: Tea, Bakery, Desserts, Seafood,
        Average Cost: 97, Aggregate Rating: 3.1, Hilton Head",
        "attraction": "Coastal Discovery Museum, Hilton Head",
        "lunch": "Name: Ashirbad, Cuisines: Pizza, Bakery, Mediterranean, Seafood,
        Average Cost: 50, Aggregate Rating: 3.6, Hilton Head",
        "dinner": "Name: Dhaba Ambarsariya, Cuisines: French, BBQ, Mediterranean,
        Seafood, Average Cost: 47, Aggregate Rating: 3.3, Hilton Head",
        "accommodation": "Name: Rustic room in renovated Bushwick apartment, room
        type: Private room, price: 309.0, minimum nights: 1.0, review rate number:
        2.0, house_rules: No smoking, maximum occupancy: 1, Hilton Head"
    },
    {
        "days": 3,
        "current_city": "from Hilton Head to Charlotte",
        "transportation": "Flight Number: F4056985, from Hilton Head to Charlotte,
        Departure Time: 20:07, Arrival Time: 21:31, Cost: 230",
        "breakfast": "Name: Mr. Brown, Cuisines: Desserts, Italian, BBQ, Fast Food,
        Cafe, Average Cost: 23, Aggregate Rating: 3.7, Hilton Head",
        "attraction": "Harbour Town Lighthouse, Hilton Head",
        "lunch": "Name: MR.D - Deliciousness Delivered, Cuisines: Tea, French,
        Mexican, Fast Food, Average Cost: 18, Aggregate Rating: 3.5, Hilton Head",
        "dinner": "-",
        "accommodation": "-"
    }
]
```

Table 19: Generic plan of the personal query created from val-98 query by Llama-3-8B-Instruct with the direct method.

```
Personal Query with pet's name: "Could you devise a 5-day travel itinerary for a group of 4, commencing in Bloomington and roaming in two cities in Florida from March 13th to March 17th, 2022? Our budget is set at $15,900. We require accommodations to be Luna friendly."
```


```
[
    {
        "days": 1,
        "current_city": "from Bloomington to Orlando",
        "transportation": "Self-driving from Bloomington to Orlando, duration: 13 hours 55 mins, distance: 1,534 km, cost: 76",
        "breakfast": "-",
        "attraction": "-",
        "lunch": "-",
        "dinner": "-",
        "accommodation": "Name: BKLYN Brownstone- Glam Getaway!, room type: Entire home/apt, price: 287.0, minimum nights: 3.0, review rate number: 3.0, house_rules: No pets & No visitors, maximum occupancy: 2, Orlando"
    },
    {
        "days": 2,
        "current_city": "Orlando",
        "transportation": "-",
        "breakfast": "-",
        "attraction": "SeaWorld Orlando, Orlando; Universal Orlando Resort, Orlando;",
        "lunch": "Name: AB's Absolute Barbecues, Cuisines: Desserts, Tea, Pizza, Italian, Fast Food, American, Average Cost: 55, Aggregate Rating: 4.8, Orlando",
        "dinner": "Name: Clocked, Cuisines: Tea, Italian, Mexican, BBQ, Cafe, Average Cost: 49, Aggregate Rating: 4.2, Orlando",
        "accommodation": "Name: BKLYN Brownstone- Glam Getaway!, room type: Entire home/apt, price: 287.0, minimum nights: 3.0, review rate number: 3.0, house_rules: No pets & No visitors, maximum occupancy: 2, Orlando"
    },
    {
        "days": 3,
        "current_city": "Orlando",
        "transportation": "-",
        "breakfast": "-",
        "attraction": "The Wheel at ICON Park, Orlando; Harry P Leu Gardens, Orlando;",
        "lunch": "Name: Anjlika, Cuisines: Pizza, Italian, Fast Food, Cafe, Mediterranean, Seafood, Average Cost: 56, Aggregate Rating: 3.1, Orlando",
        "dinner": "Name: Spices & Sauces, Cuisines: Italian, Mexican, Bakery, Cafe, Seafood, Average Cost: 28, Aggregate Rating: 3.0, Orlando",
        "accommodation": "Name: BKLYN Brownstone- Glam Getaway!, room type: Entire home/apt, price: 287.0, minimum nights: 3.0, review rate number: 3.0, house_rules: No pets & No visitors, maximum occupancy: 2, Orlando"
    },
    {
        "days": 4,
        "current_city": "from Orlando to Miami",
        "transportation": "Flight Number: F3563839, from Orlando to Miami, Departure Time: 21:15, Arrival Time: 22:28, ActualElapsedTime: 1 hours 13 minutes, FlightDate: 2022-03-15, OriginCityName: Orlando, DestCityName: Miami, Distance: 192.0",
        "breakfast": "-",
```

```
        "attraction": "-",
        "lunch": "-",
        "dinner": "-",
        "accommodation": "Name: Comfortable shared apartmant by Times Square, room
        type: Shared room, price: 924.0, minimum nights: 1.0, review rate number:
        2.0, house_rules: No smoking & No children under 10, maximum occupancy: 1,
        Miami"
    },
    {
        "days": 5,
        "current_city": "Miami",
        "transportation": "-",
        "breakfast": "-",
        "attraction": "Jungle Island, Miami; Vizcaya Museum & Gardens, Miami;",
        "lunch": "Name: Parrot's, Cuisines: Chinese, French, BBQ, Fast Food, Average
         Cost: 22, Aggregate Rating: 3.4, Miami",
        "dinner": "Name: Clocked, Cuisines: Tea, Italian, Mexican, BBQ, Cafe,
        Average Cost: 49, Aggregate Rating: 4.2, Miami",
        "accommodation": "Name: Comfortable shared apartmant by Times Square, room
        type: Shared room, price: 924.0, minimum nights: 1.0, review rate number:
        2.0, house_rules: No smoking & No children under 10, maximum occupancy: 1,
        Miami"
    }
]
```

Table 20: Personal plan of val-98 using personal queries by Llama-3-8B-Instruct with Direct method for User Model 9

```
Personal Query with pet's name: "Could you devise a 5-day travel itinerary for a
group of 4, commencing in Bloomington and roaming in two cities in Florida from
March 13th to March 17th, 2022? Our budget is set at $15,900. We require
accommodations to be Luna friendly."
```


```
[
    {
        "days": 1,
        "current_city": "from Bloomington to Orlando",
        "transportation": "Flight Number: F3563219, from Bloomington to Orlando,
        Departure Time: 15:35, Arrival Time: 19:17, Cost: 302",
        "breakfast": "-",
        "attraction": "SeaWorld Orlando, Orlando; The Wheel at ICON Park, Orlando;
        Fun Spot America Theme Parks, Orlando;",
        "lunch": "-",
        "dinner": "Name: Fuji Japanese Steakhouse, Cuisines: Tea, Bakery, Desserts,
        Average Cost: 75, Aggregate Rating: 3.8, Orlando; restaurants_664, Orlando",
        "accommodation": "Name: Private room in Jackson Heights Apartment 2+, room
        type: Private room, price: 1081.0, minimum nights: 1.0, review rate number:
        1.0, house_rules: No parties, maximum occupancy: 1, Orlando"
    },
    {
        "days": 2,
        "current_city": "Orlando",
        "transportation": "-",
        "breakfast": "Name: Fuji Japanese Steakhouse, Cuisines: Tea, Bakery,
        Desserts, Average Cost: 75, Aggregate Rating: 3.8, Orlando",
        "attraction": "Harry P Leu Gardens, Orlando; Universal Orlando Resort,
        Orlando",
        "lunch": "Name: The Tandoori Times, Cuisines: Tea, Cafe, BBQ, Italian,
        Average Cost: 84, Aggregate Rating: 2.7, Orlando",
        "dinner": "Name: Chaayos, Cuisines: Tea, Cafe, Mexican, Average Cost: 73,
        Aggregate Rating: 3.9, Orlando",
        "accommodation": "Name: Private room in Jackson Heights Apartment 2+, room
        type: Private room, price: 1081.0, minimum nights: 1.0, review rate number:
        1.0, house_rules: No parties, maximum occupancy: 1, Orlando"
    },
    {
        "days": 3,
        "current_city": "from Orlando to Miami",
        "transportation": "Flight Number: F3563839, from Orlando to Miami, Departure
         Time: 21:15, Arrival Time: 22:28, Cost: 70",
        "breakfast": "-",
        "attraction": "Jungle Island, Miami; P\u00c3\u00a8rez Art Museum Miami,
        Miami",
        "lunch": "Name: Clocked, Cuisines: Tea, Italian, Mexican, BBQ, Cafe, Average
         Cost: 49, Aggregate Rating: 4.2, Miami; restaurants_175, Miami",
        "dinner": "-",
        "accommodation": "Name: Simple and clean bedroom with good view and light,
        room type: Private room, price: 669.0, minimum nights: 1.0, review rate
        number: 5.0, house_rules: No children under 10, maximum occupancy: 2, Miami"
    },
    {
        "days": 4,
        "current_city": "Miami",
        "transportation": "-",
        "breakfast": "Name: Clocked, Cuisines: Tea, Italian, Mexican, BBQ, Cafe,
        Average Cost: 49, Aggregate Rating: 4.2, Miami",
        "attraction": "Vizcaya Museum & Gardens, Miami; Miami Children's Museum,
        Miami",
```

```
            "lunch": "Name: Papouli's Mediterranean Cafe & Market, Cuisines: Tea,
            Desserts, Average Cost: 26, Aggregate Rating: 4.2, Miami",
            "dinner": "Name: Parrot's, Cuisines: Chinese, French, BBQ, Fast Food,
            Average Cost: 22, Aggregate Rating: 3.4, Miami",
            "accommodation": "Name: Simple and clean bedroom with good view and light,
            room type: Private room, price: 669.0, minimum nights: 1.0, review rate
            number: 5.0, house_rules: No children under 10, maximum occupancy: 2, Miami"
        },
        {
            "days": 5,
            "current_city": "from Miami to Bloomington",
            "transportation": "Self-driving from Miami to Bloomington, which takes
            approximately 17 hours 1 min and costs 94",
            "breakfast": "-",
            "attraction": "-",
            "lunch": "-",
            "dinner": "-",
            "accommodation": "-"
        }
]
```

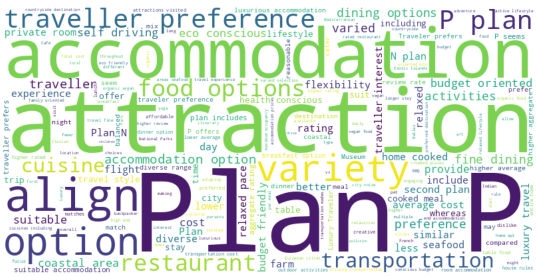



Figure 4: Word cloud of reasoning from judge Llama-3-8B-Instruct for validation queries on Direct method. Plans are generated by Llama-3-8B-Instruct.

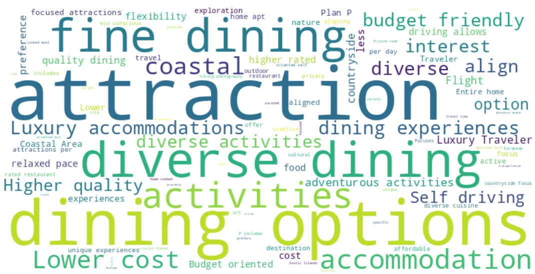


Figure 5: Word cloud of reasoning from judge Gemma2-9B-Instruct for validation queries on Direct method. Plans are generated by Llama-3-8B-Instruct.

Figure 6: Word cloud of reasoning from judge Llama-3-8B-Instruct for validation queries on Direct method. Plans are generated by GPT-3.5-Turbo.

Figure 7: Word cloud of reasoning from judge Gemma2-9B-Instruct for validation queries on Direct method. Plans are generated by GPT-3.5-Turbo.

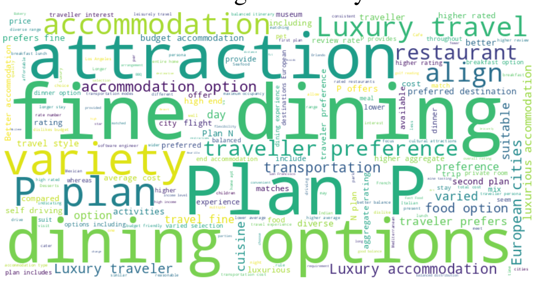


Figure 8: Word cloud of reasoning from judge Llama-3-8B-Instruct for test queries on Direct method. Plans are generated by Llama-3-8B-Instruct.

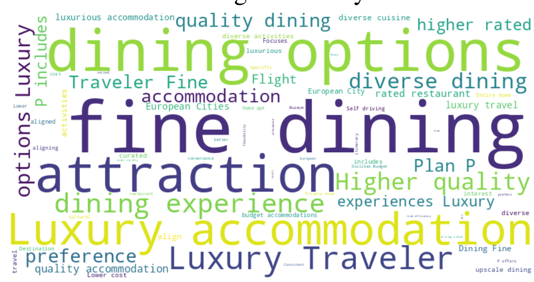


Figure 9: Word cloud of reasoning from judge Gemma2-9B-Instruct for test queries on Direct method. Plans are generated by Llama-3-8B-Instruct.

Table 21: Example to illustrate the issues of context window limitations and hallucinations encountered using the ReAct prompting strategy with the Llama-3-8B-instruct LLM agent

Thought prompt for LLM

```
You are a proficient planner. Based on the provided query and information, please
give me concise thoughts on how to solve this task... Don't provide any action in
this step.

Please think about making plan for a single day only at each step.

[Anonymized one-shot example]

Given information: { reference information }

Query: Can you devise a week-long travel plan for a solo traveler? The trip takes
off from Columbus and involves visiting 3 distinct cities in Texas from March 1st to
 March 7th, 2022. The budget for this venture is set at $4,200.
Thought 1:
```

LLM Response

```
Based on the provided information, I'll outline a week-long travel plan for a solo
traveler from Columbus to Texas, considering the budget of $4,200.
    ** Day 1: March 1st, 2022 - Columbus to Dallas **
        * Flight from Columbus to Dallas: Book flight F3666359 (Price: $208) or
        F3712939 (Price: $374) for a total cost of $208 or $374.
        * Accommodation in Dallas: Book accommodation "accommodations_957" (Price:
        $746) for a total cost of $746.
    ...
    ** Day 7: March 7th, 2022 - Houston to Columbus **
        * Flight from Houston to Columbus: Book flight F3997947 (Price: $259) or
        F4042657 (Price: $428) for a total cost of $259 or $428.
        * Transportation: Self-driving from Houston to Columbus (Duration: 17 hours
        20 mins, Distance: 1,858 km, Cost: $92) or take a taxi (Duration: 17 hours
        20 mins, Distance: 1,858 km, Cost: $1,858).
    ...
This plan should provide a comfortable and enjoyable trip for the solo traveler,
staying within the budget of $4,200. However, please note that the estimated costs
are based on the provided information and may vary depending on actual prices and
availability.
```

Action prompt for LLM

```
You are a proficient planner with a keen understanding of preferences and styles...
The 'Action' phase should consist of planning, that can be only one of two types:
- CostEnquiry[Sub Plan]: ...
- Finish[Final Plan]: ...
...
Query: Can you devise a week-long travel plan for a solo traveler? The trip takes
off from Columbus and involves visiting 3 distinct cities in Texas from March 1st to
 March 7th, 2022. The budget for this venture is set at $4,200.
Thought 1: Based on the provided information, ...
Action 1:

Now, based on your reasoning in the Thought phase, please proceed with only one
necessary action from CostEnquiry or Finish.
```

LLM Response

```
Max Token Length Exceeded
```

Observation

```
Action None is not supported
...
```

Table 22: Prompt example for preference evaluation on the generic and personalized plans

```
You are a evaluator for travel plans. You are provided the persona of the traveller
and two travel plans named N and P. You should pick the more suitable plan based on
the traveller persona. Output only N or P for choosing between the two plans. The
persona of the traveller is {persona}. The N plan is {n_plan}. The P plan is {p_plan
}.
```

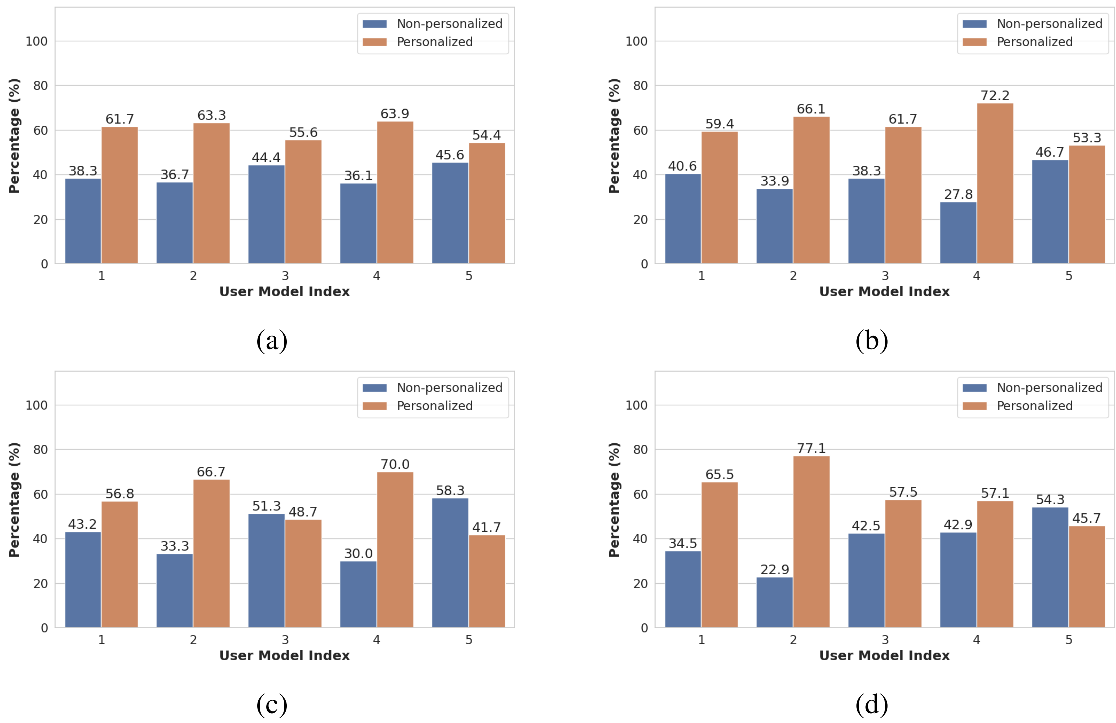


Figure 10: Preference rates for plans generated by Llama-3-8B-Instruct using various methods for validation split on 20 user models, with Llama-3-8B-Instruct as judge, (a) Direct, (b) CoT, (c) ReAct, (d) Reflexion.

Table 23: Comparison between original and improved prompting for two strategies. The performance indicators (%) were calculated with Llama-3-8B-instruct LLM agent and two planning strategies on the TravelPlanner validation set. The Personal plan results correspond to a single randomly chosen user model. The best outcomes are in bold.

Planning strategy	Generic plans						Personal plans					
	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro			Micro	Macro	Micro	Macro	
Original prompts												
Direct	100	71.5	5.5	12.4	3.3	0	**100**	67.6	6.1	7.5	5.6	1.7
CoT	100	64.2	1.1	8.7	5.0	0.6	99.44	61.1	0.0	5.9	3.9	0
Improved prompts												
Direct	**100**	**76.53**	**16.11**	**31.67**	**8.33**	1.67	**98.89**	**72.36**	**13.33**	**16.19**	**6.67**	**1.67**
CoT	98.89	69.65	8.33	16.43	5.00	**2.22**	98.89	67.50	4.44	9.76	4.44	1.11

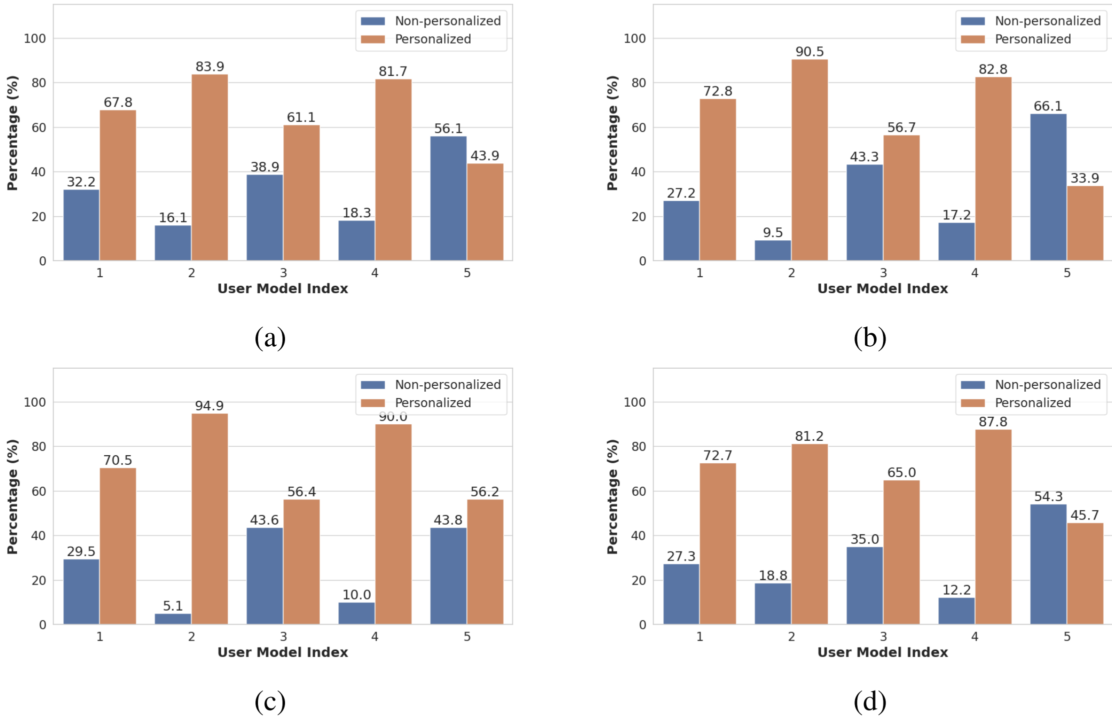


Figure 11: Preference rates for plans generated by Llama-3-8B-Instruct using various methods for validation split on 20 user models, with Gemma2-9B-Instruct as judge, (a) Direct, (b) CoT, (c) ReAct, (d) Reflexion.

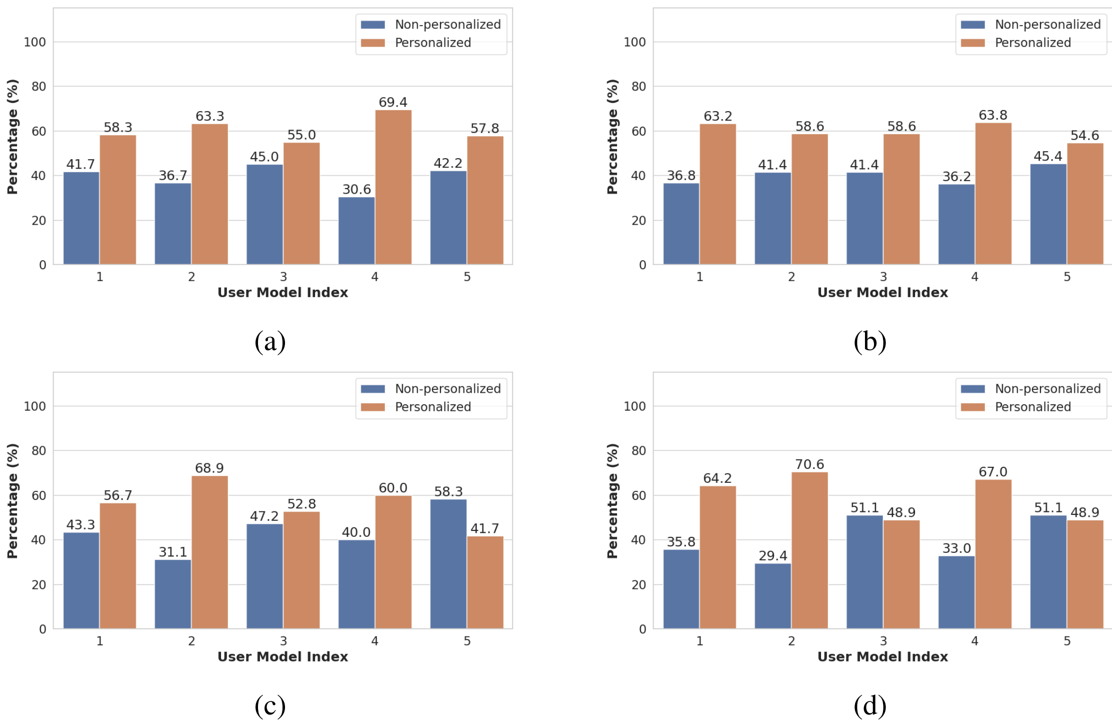


Figure 12: Preference rates for plans generated by GPT-3.5-Turbo using various methods for validation split on 20 user models, with Llama-3-8B-Instruct as judge, (a) Direct, (b) CoT, (c) ReAct, (d) Reflexion.

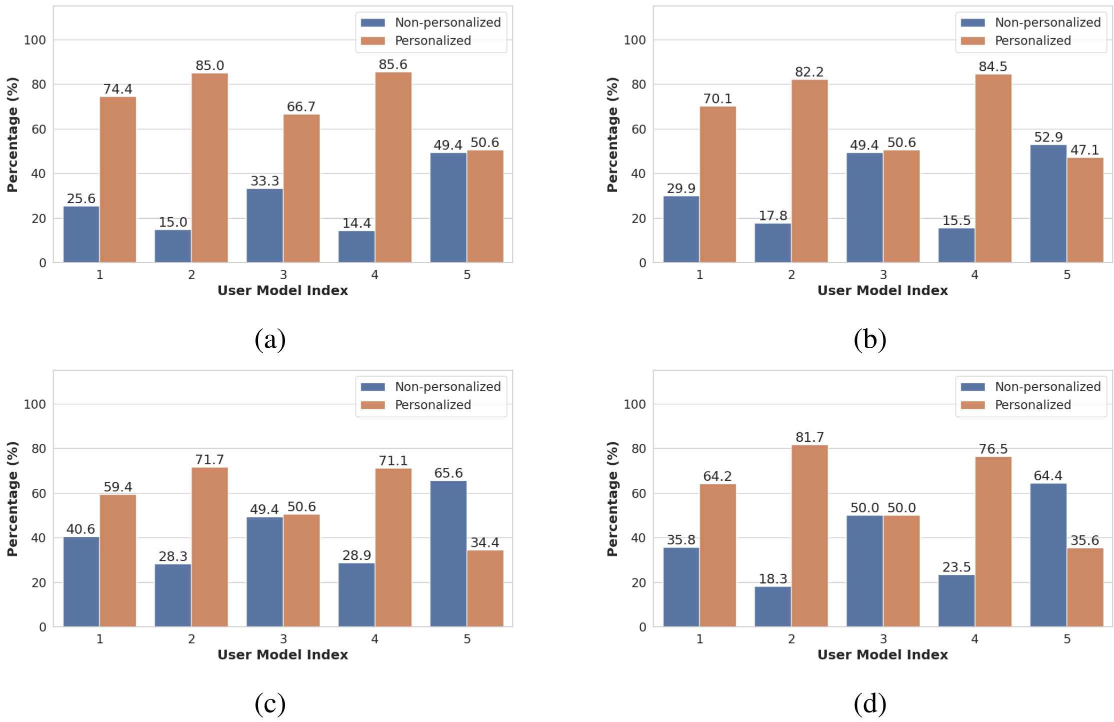


Figure 13: Preference rates for plans generated by GPT-3.5-Turbo using various methods for validation split on 20 user models, with Gemma2-9B-Instruct as judge, (a) Direct, (b) CoT, (c) ReAct, (d) Reflexion.

Table 24: Performance Indicators (%) of GPT-3.5-Turbo LLM Agent with all Planning Strategies on TravelPlanner prompt on validation split compared to improved prompts in TravelPlanner+. The best values are in bold.

Strategy	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro	
TravelPlanner						
Direct	**100**	60.2	**4.4**	11.0	2.8	0
CoT	**100**	66.3	3.3	11.9	5.0	0
ReAct	82.2	47.6	3.9	11.4	6.7	0.6
Reflexion	93.9	53.8	2.8	11.0	2.8	0
TravelPlanner+						
Direct	**100**	67.2	3.3	20.2	5.0	0
CoT	**100**	**66.9**	3.3	**20.9**	**9.4**	**1.11**
ReAct	**100**	64.4	2.2	9.3	2.8	0
Reflexion	**100**	63.5	0.6	3.6	1.1	0

Table 25: Performance indicators (%) with Qwen-1.5-7B-chat LLM agent for all planning strategies on the TravelPlanner validation set. The Personal plan results are averaged over five different user models.

Planning strategy	Generic plans						Personal plans					
	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate	Delivery Rate	Commonsense Pass Rate		Hard Constraint Pass Rate		Final Pass Rate
		Micro	Macro	Micro	Macro			Micro	Macro	Micro	Macro	
Qwen-1.5-7B-chat												
Direct	100	52.71	0.56	5.00	0	0	100	49.81	0.45	3.64	0.11	0
CoT	100	48.89	0.56	3.00	0	0	99.88	50.10	0.22	4.09	0.22	0
ReAct	100	55.97	0	3.33	0	0	100	57.29	0.11	10.5	0	0
Reflexion	100	52.92	0	2.50	0.56	0	100	54.68	0	2.67	0	0