PersonaLens: A Benchmark for Personalization Evaluation in Conversational AI Assistants

Zheng Zhao[1], Clara Vania[2], Subhradeep Kayal[2], Naila Khan[2], Shay B. Cohen[1], Emine Yilmaz[2,3]

[1]University of Edinburgh [2]Amazon

[3]University College London

Overview

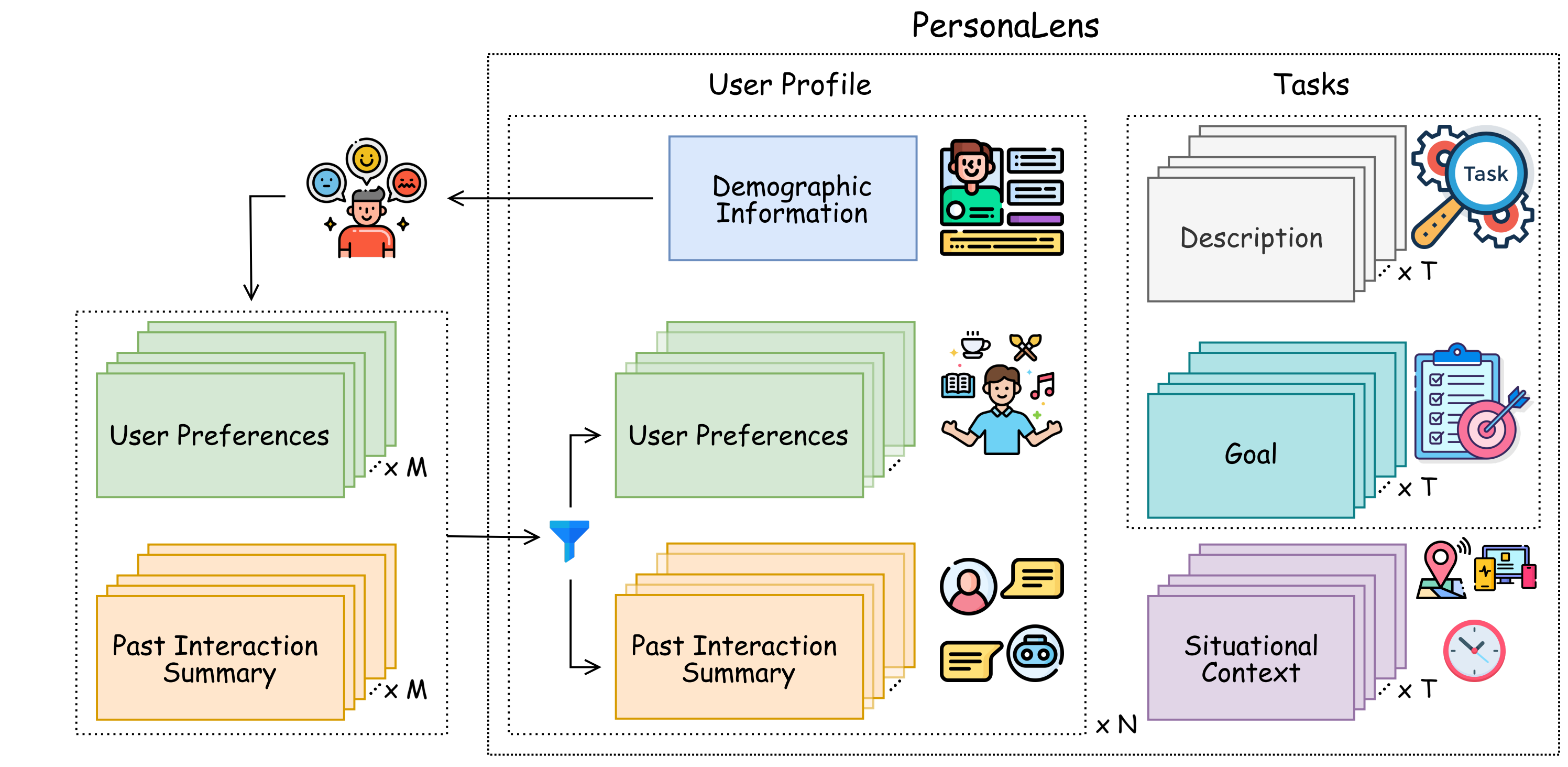
- This work introduces **PersonaLens**, a comprehensive benchmark designed to evaluate the **personalization capabilities** of large language models (LLMs) within task-oriented conversational AI assistants.
- The benchmark features **1,500 diverse user profiles** with rich preferences and interaction histories, alongside two specialized LLM agents: a **User Agent** for realistic dialogue simulation and a **Judge Agent** for automated evaluation.
- Our study, using PersonaLens to benchmark leading LLMs, reveals two key findings: current models exhibit **limited personalization**, especially in complex multi-domain scenarios, and more critically, **past interaction history** is the most important factor for tailoring responses, far outweighing static user data.

Motivation & Contribution

Large Language Models (LLMs) have revolutionized conversational AI. However, evaluating how well they **personalize** responses to individual users during task-oriented dialogues remains a major challenge. Existing benchmarks are often limited to:

- Chit-chat scenarios (e.g., PersonaChat)
- Narrow, single-domain tasks

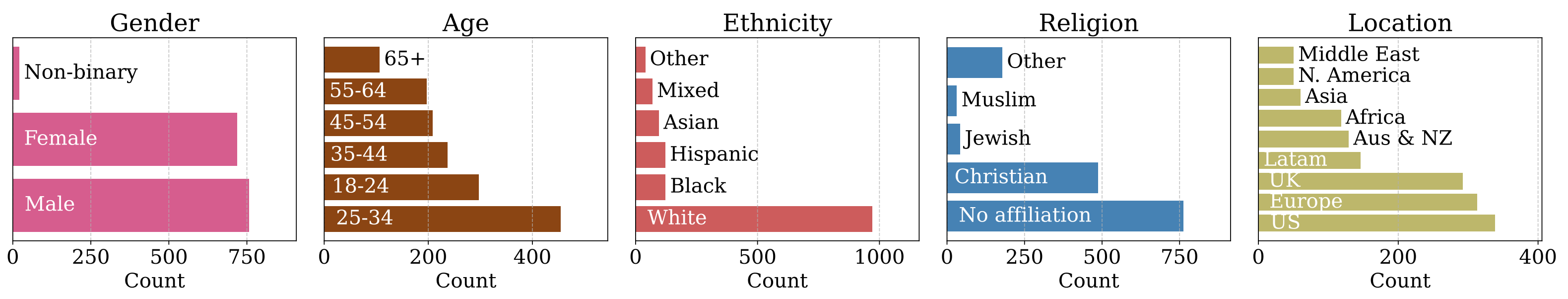
The PersonaLens Benchmark



PersonaLens is a large-scale benchmark for evaluating the **personalization** of AI assistants in task-oriented dialogues, built on three core components:

- **1,500 Diverse User Profiles.** Each profile of a user contains:
  - **Demographics:** Attributes like age, gender, and ethnicity from real users across 75 countries.
  - **User Preferences:** Detailed categorical and open-ended preferences across various domains.
  - **Interaction History:** Natural language summaries of past user-assistant exchanges.
- **111 Task-Oriented Scenarios:**
  - **Scope:** 111 tasks across 20 domains, composed of 86 single-domain and 25 multi-domain tasks.
  - **Dynamic Features:** Each task is enriched with a **situational context** and personalized using a **binary mask** to filter by user interest.
  - **Resulting Scale:** This generates 122,133 unique user-task scenarios (98,115 single-domain & 24,018 multi-domain).
- **Two LLM-Powered Agents.** A **User Agent** that simulates realistic user behavior based on a given profile and task, and a **Judge Agent** that systematically evaluates the assistant's dialogue for personalization, quality, and task success.

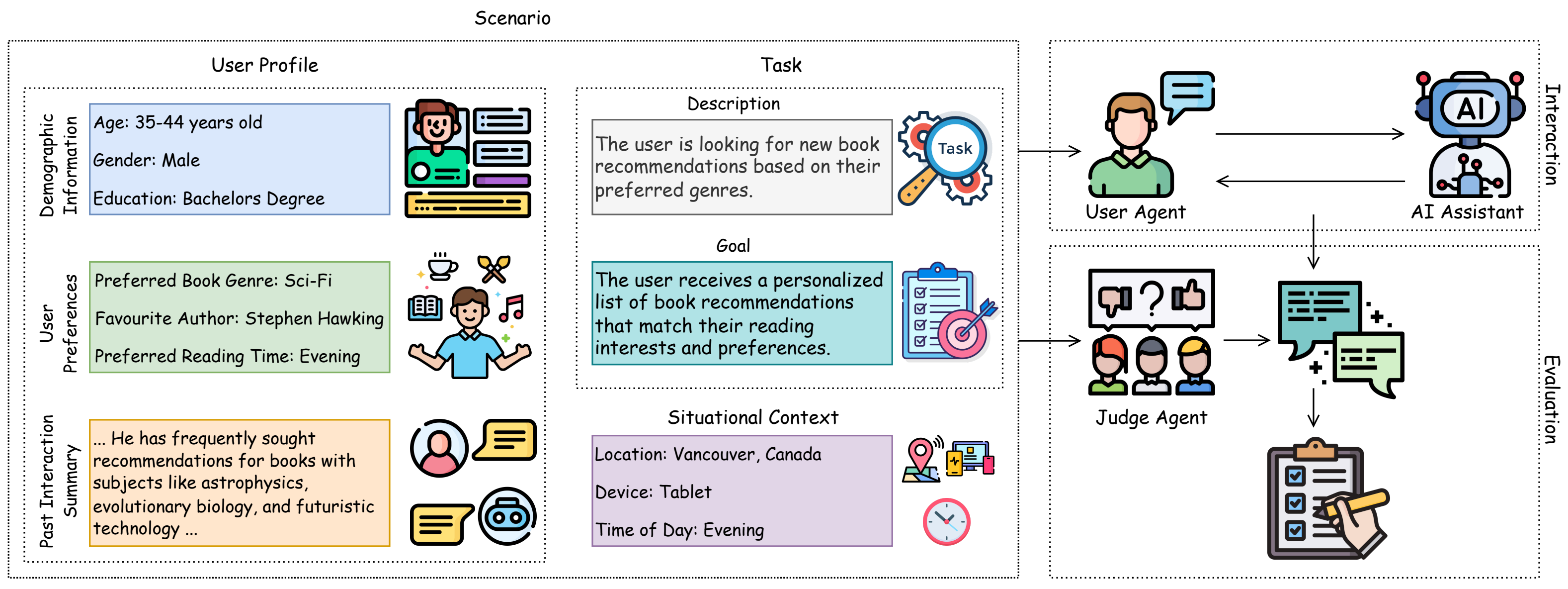
Our benchmark's demographic diversity is grounded in the PRISM Alignment dataset, resulting in the user distribution shown below:



Materials

(a) Paper (b) Data/Code (c) Connect with Me

Using the Benchmark


Scenario
User Profile
Demographic Information
Age: 35-44 years old
Gender: Male
Education: Bachelors Degree
User Preferences
Preferred Book Genre: Sci-Fi
Favourite Author: Stephen Hawking
Preferred Reading Time: Evening
Past Interaction Summary
... He has frequently sought recommendations for books with subjects like astrophysics, evolutionary biology, and futuristic technology ...
Task
Description
The user is looking for new book recommendations based on their preferred genres.
Task
Goal
The user receives a personalized list of book recommendations that match their reading interests and preferences.
Situational Context
Location: Vancouver, Canada
Device: Tablet
Time of Day: Evening
User Agent
AI
AI Assistant
Interaction
Judge Agent
Evaluation

- The benchmark provides a complete **user-task scenario** to the User Agent, including the user profile, task specification, and situational context.
- The **User Agent** interacts with the AI Assistant being evaluated, simulating a real user and generating a multi-turn dialogue.
- The **Judge Agent** then analyzes the entire dialogue based on the original user profile and task scenario.
- Finally, the Judge Agent provides **feedback** on the assistant's performance, assessing personalization, task success, and overall quality.

Experimental Setup

- **Models Evaluated:** 7 leading LLM assistants across 4 model families.
- **Experimental Scale:** For computational feasibility, experiments were run on a sampled subset of **50 user profiles**, generating **3,283 single-domain** and **813 multi-domain** dialogues for analysis.
- **Key Evaluation Metrics:**
  - **Task Completion Rate (TCR):** The percentage of tasks successfully completed.
  - **Personalization (P):** How well responses are tailored to the user profile (1-4 scale).
  - **Naturalness & Coherence:** Dialogue quality rated for human-likeness and consistency (1-5 scale).

Main Results

Assistant Model	T_{SD}				T_{MD}			
	TCR↑	P↑	Nat.↑	Coh.↑	TCR↑	P↑	Nat.↑	Coh.↑
Claude 3 Haiku	95.95%	2.20	3.77	4.62	75.65%	1.98	3.78	4.66
Claude 3.5 Haiku	91.53%	2.32	4.01	4.86	70.85%	2.18	4.08	4.88
Claude 3 Sonnet	95.98%	2.13	3.86	4.71	77.49%	2.01	3.84	4.79
Llama 3.1 8B Instruct	89.55%	2.14	3.90	4.68	77.00%	2.03	3.64	4.33
Llama 3.1 70B Instruct	90.80%	2.21	4.11	4.86	83.03%	2.22	4.02	4.89
Mistral 7B Instruct	88.52%	1.93	3.49	4.38	74.54%	1.86	3.18	4.07
Mixtral 8x7B Instruct	91.38%	2.04	3.88	4.76	78.35%	2.00	3.77	4.67

Table 1. Evaluation results of assistant models on T_{SD} and T_{MD} tasks. TCR: task completion rate, P: personalization. Naturalness (Nat.) and Coherence (Coh.) here refer to the assistant's responses. ↑ denotes higher is better.

Analysis

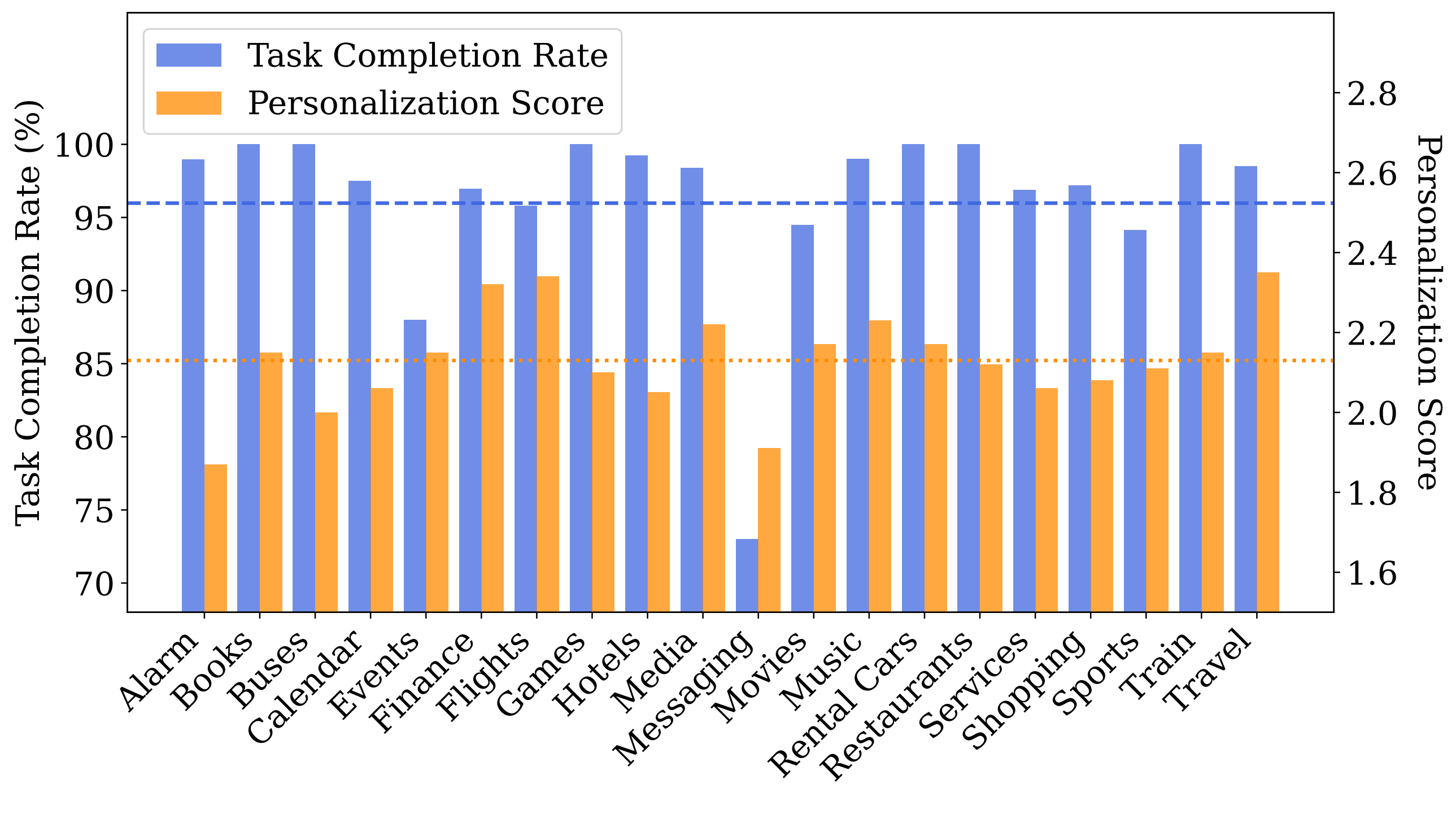


Figure 2. Evaluation results of the assistant (Claude 3 Sonnet) by domain. The dashed line is the average performance over all domains.

Setting	T_{SD}		T_{MD}	
	TCR↑	P↑	TCR↑	P↑
Vanilla	92.93%	2.16	75.40%	2.08
Base	95.98%	2.13	77.49%	2.01
Base + D	95.52%	2.16	77.86%	2.05
Base + I	**96.83%**	**2.59**	81.30%	**2.32**
Base + S	95.74%	2.20	77.61%	2.06
Base + all	96.31%	2.57	**82.66%**	2.31

Table 2. Ablation studies on the effect of varying levels of instruction and additional information provided to the assistant (Claude 3 Sonnet). "Vanilla" uses minimal instructions, while "Base" uses instructions emphasizing personalization. D: demographic information; I: past interaction summary; S: situational context. "all" means D + I + S. TCR: Task completion rate, P: Personalization. ↑ denotes higher is better.

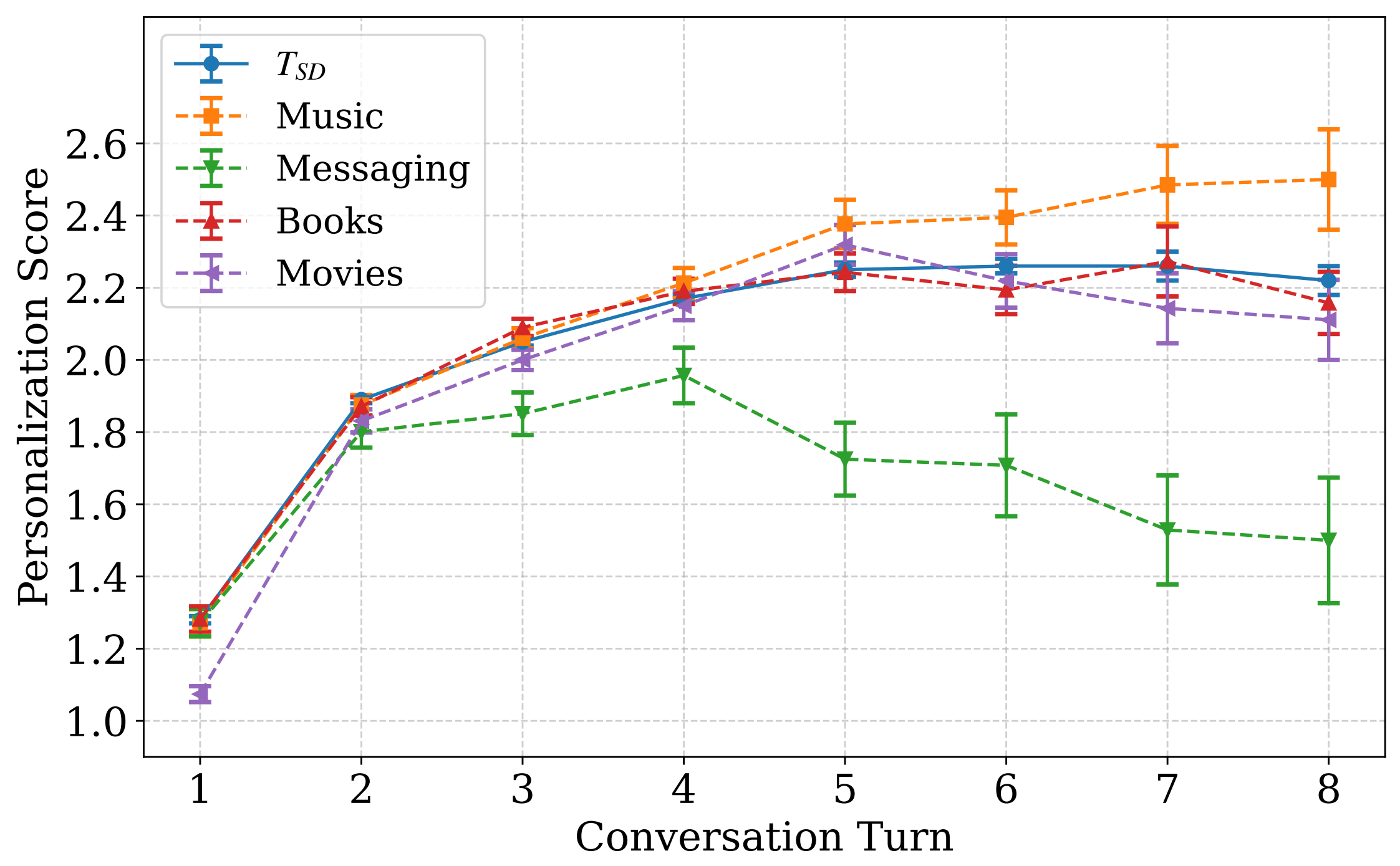


Figure 3. Results on turn-level personalization for the assistant (Claude 3 Sonnet).