

# IP-Dialog : Evaluating Implicit Personalization in Dialogue Systems with Synthetic Data

Bo Peng<sup>1,2,3\*</sup>, Zhiheng Wang<sup>1,2\*</sup>, Heyang Gong<sup>4</sup>, Chaochao Lu<sup>1,3†</sup>

<sup>1</sup>Shanghai Artificial Intelligence Laboratory, <sup>2</sup>Shanghai Jiao Tong University

<sup>3</sup>Shanghai Innovation Institute <sup>4</sup>Sicore Ladder Tech Co. Ltd.

peng\_bo2019@sjtu.edu.cn, wangzhiheng@pjlab.org.cn,

zj3712@gmail.com, luchaochao@pjlab.org.cn

## Abstract

In modern dialogue systems, the ability to implicitly infer user backgrounds from conversations and leverage this information for personalized assistance is crucial. However, the scarcity of high-quality data remains a fundamental challenge to evaluating and improving this capability. Traditional dataset construction methods are labor-intensive, resource-demanding, and raise privacy concerns. To address these issues, we propose a novel approach for automatic synthetic data generation and introduce the **Implicit Personalized Dialogue (IP-Dialog)** benchmark along with a training dataset, covering 10 tasks and 12 user attribute types. Additionally, we develop a systematic evaluation framework with four metrics to assess both attribute awareness and reasoning capabilities. We further propose five causal graphs to elucidate model reasoning pathways during implicit personalization. Extensive experiments yield insightful observations and prove the reliability of our dataset. Our dataset and code are available at <https://github.com/OpenCausaLab/IP-Dialog>.

## 1 Introduction

Implicit personalization (IP) (Flek, 2020; Raharjana et al., 2021; Jin et al., 2024), which involves tailoring responses based on inferred user characteristics without explicit user profiles, is crucial for enhancing the user experience in various AI-driven systems, including conversational agents (Anantha et al., 2021; Singhal et al., 2023; Zhuang et al., 2023), recommendation systems (Wang et al., 2023a), and personalized content delivery (Qian et al., 2024). In human-AI dialogues, user identities are implicitly embedded in the context of their inputs. These latent identities are vital in determining user preferences and shaping the expected AI

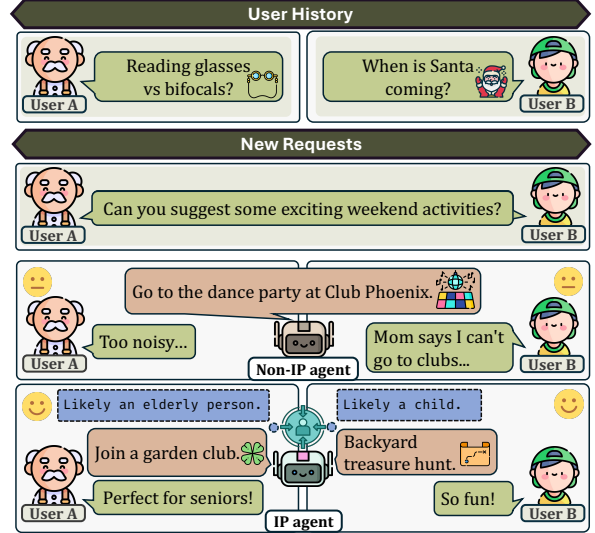


Figure 1: A comparative example of an AI agent with implicit personalization capability (IP agent) and one without (non-IP agent). The IP agent infers implicit user identities from dialogue history and generates customized responses accordingly.

responses (Flek, 2020; Raharjana et al., 2021). Figure 1 shows an example where AI agents answer questions based on user histories. An AI agent without IP capability may provide unsuitable suggestions, as it fails to infer user identities from the history. In contrast, an IP-capable agent can deliver personalized answers by recognizing users' latent identities (e.g., an elderly person or a child). Such implicit personalization enables AI systems to provide more appropriate and engaging responses through a user-friendly approach.

However, no evaluation benchmarks or standards are available for IP, as publishing detailed user information causes privacy violation risks (Carlini et al., 2021, 2023). Moreover, conventional manually labeled dataset construction approaches are prohibitively expensive and time-consuming. Considering the success of synthetic data (Xu et al., 2024; Lou et al., 2024; Yuxhymenko et al., 2024b; Zheng et al., 2023), we decide to utilize

\*Equal contribution.

†Corresponding author.

this advancement and propose an automated data generation pipeline powered by state-of-the-art LLMs. With this pipeline, we establish the **Implicit Personalized Dialogue (IP-Dialog)** benchmark.

Our benchmark covers three carefully designed scenarios, encompassing 10 tasks with four distinct answer formats. We characterize users through 12 key attribute types (e.g., age, profession). Each benchmark item consists of a user history for attribute inference and a user question that requires the model to incorporate the inferred attributes into its response. The user questions are generated through a multi-stage process: starting with 10 to 15 manually curated domains (e.g., sports, education) per task, generating 10 model-produced subjects per domain, and finally creating 10 user questions per subject that span diverse user attribute combinations. The user history is constructed iteratively, with each dialogue turn refined to reflect a single user attribute. The resulting dataset is divided into a training set (10,790 samples) and the IP-Dialog benchmark (1,000 samples).

To systematically evaluate the IP capabilities of models, we establish a comprehensive evaluation framework comprising four primary metrics: two measuring attribute awareness and two evaluating attribute-based reasoning abilities. Furthermore, we propose five causal graphs to model how LLMs reason within the IP-Dialog task. These graphs range from a basic approach that disregards user attributes to more sophisticated reasoning pathways involving hidden attribute prediction and relevant attribute identification. Finally, we conduct extensive experiments across six models, yielding the following key findings:

1. Models that excel at identifying relevant attribute types also demonstrate high accuracy in predicting the correct attribute values.
2. Claude-3.5-Sonnet achieves the best performance across all metrics. Both Claude-3.5-Sonnet and GPT-4o have outperformed humans in solving IP tasks.
3. Tasks in the **behavior analysis** scenario, such as action prediction and preference inference, present the greatest challenge due to their dependence on complex psychological factors.
4. The most effective reasoning pathway is **Type-Guided** which begins with inferring *related attribute types*, followed by guessing *related attributes* and finally providing the *response*. **TaskRelated** serves as a viable alternative

by directly inferring *related attributes* before responding. Their high performance is mainly due to the precise and efficient attribute-related consideration process. Models with stronger IP capabilities show more resilience to variations in reasoning pathways.

5. Supervised fine-tuning (SFT) significantly enhances the IP capability of Llama-3.1-8B-Instruct beyond all other models. Models after SFT adapt well to unseen tasks with familiar answer formats but struggle with new formats. Moreover, SFT on a single reasoning pathway improves performance across other pathways.

Our contributions are summarized as follows:

- We design an efficient and highly controllable synthetic data methodology, providing solutions to data scarcity, privacy risks and evaluation challenges across various AI applications.
- We introduce the IP-Dialog benchmark and the corresponding evaluation framework. To our knowledge, we are the first to evaluate the IP capabilities of LLMs in dialogue systems.
- We explore the impact of reasoning pathways on model performance in IP through five hypothesized causal graphs.
- Extensive experiments yield insightful observations and five key findings.

## 2 Design of IP-Dialog

Current AI-human dialogues can be conceptualized as consisting of a user historical dialogue (user history,  $H$ ) and the current user request (user question,  $Q$ ). The user history encapsulates the user’s *hidden attributes*  $A$ , which are not explicitly stated but can be inferred from past interactions. Implicit personalization (IP) in dialogues can be defined as a two-step process: first, inferring the *related attributes*  $A_s$  relevant to  $Q$  from  $H$ , and then leveraging  $A_s$  to generate personalized responses. Following this definition, we construct the IP-Dialog benchmark, where each benchmark item consists of a task name, user history  $H$ , user question  $Q$ , *related attributes*  $A_s$ , and the ground truth *response*, which includes both *analysis* and *answer* components.

### 2.1 User Attributes

To comprehensively model user diversity, we design 12 attribute types that significantly influence users’ needs, preferences, and behavior patterns: age, gender, income level, profession, residence,

Task name	Definition
Recommendation System (Guo et al., 2017; Lin et al., 2023)	
Recommendation (Rec, Open-ended)	<b>Definition:</b> Generate personalized item recommendations based on user attributes. (Guo et al., 2017) <b>Example:</b> I’m looking for high-performance sports gear that can support me during my workouts.
Ranking (Rank, Ranking)	<b>Definition:</b> Prioritize items based on user preferences and requirements. (Rendle et al., 2012) <b>Example:</b> Please rank the following online courses in order of suitability for me: A. online music course B. online painting course C. online cooking course D. online language course
Filtering (Fil, Multiple choice)	<b>Definition:</b> Select items that match specific user needs. (He et al., 2017) <b>Example:</b> Which of the following car models would appeal to my lifestyle: A. compact electric car B. rugged off-road SUV C. luxury convertible sports car D. practical minivan E. classic muscle car
Behavior Analysis (Morita and Shinoda, 1994; Jin et al., 2013)	
Prediction (Pred, Binary choice)	<b>Definition:</b> Predict whether a user will take a specific action. (Pi et al., 2019; Megargee, 1976) <b>Example:</b> Predict whether I will attend a live basketball game this weekend.
Preference Inference (PI, Multiple choice)	<b>Definition:</b> Infer user’s preference patterns by analyzing attribute-preference correlations. (Yang et al., 2014; Kelly and Teevan, 2003) <b>Example:</b> What kind of exercise environment would I enjoy more? A. Group classes B. Home workouts C. Outdoor activities D. Gym workouts
Risk Detection (RD, Open-ended)	<b>Definition:</b> Identify potential risks based on user characteristics. (Blanchard et al., 2011) <b>Example:</b> What risks do you think I might face?
Intention Inference (II, Open-ended)	<b>Definition:</b> Uncover underlying motivations from ambiguous user requests. (Kuo and Chen, 2023a) <b>Example:</b> Can you determine why I might want to explore specific athlete training routines?
Action Guide (Ajzen, 1985)	
Advice (Adv, Open-ended)	<b>Definition:</b> Provide practical solutions to user problems. (Wang and Torres, 2022; Kim et al., 2024) <b>Example:</b> I’m interested in finding a way to incorporate sports into my routine. Any suggestions?
Decision (Dec, Binary choice)	<b>Definition:</b> Make yes/no suggestions for user decisions. (Yaniv, 2004; Green and Chen, 2019) <b>Example:</b> Should I take up yoga classes to help manage stress and improve flexibility?
Convincing (Conv, Open-ended)	<b>Definition:</b> Develop persuasive arguments tailored to user characteristics. (Zeng et al., 2024; Prakken, 2006; Yoshino et al., 2018) <b>Example:</b> Convince me to try rock climbing.

Table 1: Task name (along with abbreviation and answer format), definitions, and examples of task questions.

Big Five personality traits, health status, and personal interests. These attribute types are well-established in personalization literature (Yukhymentko et al., 2024b; Petrak et al., 2024; Yang et al., 2021; Liu et al., 2023; Azevedo, 2013), striking a balance between annotation feasibility and user coverage. Appendix A.1 provides their corresponding attribute values.

## 2.2 Tasks

We categorize our 10 proposed tasks into three practical application scenarios: **Recommendation System**, **Behavior Analysis**, and **Action Guide**. To accommodate diverse task requirements, we define four distinct answer formats: **open-ended**, **ranking**, **multiple-choice**, and **binary-choice**. See Table 1 for task definitions.

## 3 Construction of IP-Dialog

Figure 2 illustrates our dataset generation pipeline. We construct a total of 11,790 items, from which we randomly sample 1,000 items to form the IP-Dialog benchmark, ensuring efficiency and cost-

effectiveness in evaluation. The remaining items constitute the training set. To ensure the dataset’s reliability, we carefully designed the generation process, containing a large amount of process checking. Detailed statistics and pseudo-code are given in Appendix B.

### 3.1 User Attributes Construction

For each attribute type, we randomly select an attribute value to form a user’s *candidate attributes*. We implement specific constraints to prevent unrealistic attribute combinations, such as assigning “retired” as a profession for a child.

### 3.2 User Question Construction

For each task, we begin by manually selecting 10-15 most common domains. Next, we prompt GPT-4o (OpenAI, 2023) to generate 10 relevant subjects for each domain based on the task description. For each subject, we provide GPT-4o with the *candidate attributes* and instruct it to generate user questions. To ensure high-quality generation, we include manually crafted examples as guidance. Once a subject or user question is generated, we

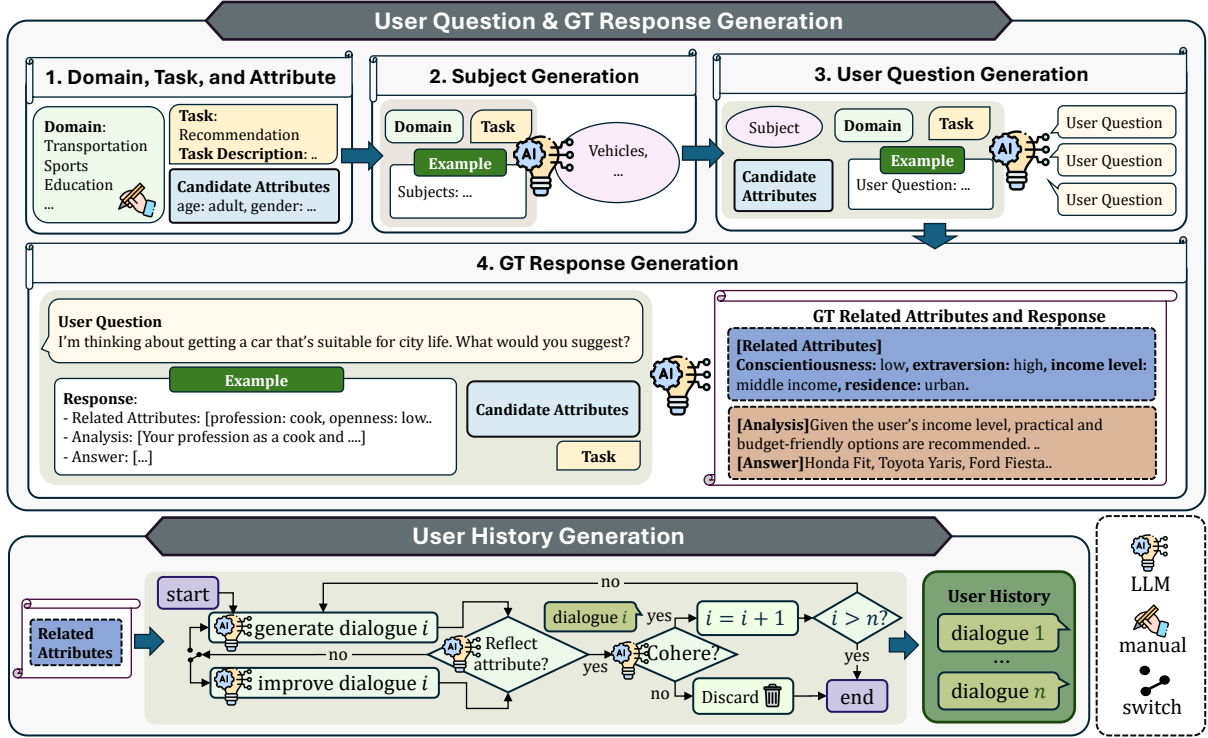


Figure 2: Construction pipeline of the IP-Dialog dataset. User questions and ground-truth (GT) responses are generated through: (1) Design domains, tasks, and attributes manually; (2) Generate subjects with LLM based on domain, task and manually designed examples; (3) Generate user questions with LLM based on subject, domain, task, manually designed examples and the candidate attributes; (4) Generate GT responses with LLM based on task, user question and candidate attributes. User history is generated based on the *related attributes* derived from (4). In each step, we generate a single attribute  $i$ . We introduce interactive checks and regeneration to ensure the attribute is reflected in the dialogue and the dialogue is coherent with all related attributes.

sample 15 items for quality check. If any fails the check, we refine the prompts and regenerate. After generating user questions, GPT-4o identifies the *related attributes* from the *candidate attributes*, performs *analysis* on how these attributes influence the user’s need, and finally generates the *answer*.

### 3.3 User History Construction

We utilize *related attributes* to generate *user history*. During construction, we find that generating history that reflects all *related attributes* in one turn directly is challenging. Therefore, we design to generate history with  $|related\ attributes|$  steps and a check-and-refine procedure. At each step, GPT-4o generates a single-round dialogue  $i$  that implicitly reflects one related attribute  $i$  from the *related attributes*. Then, GPT-4o verifies whether the generated dialogue can reflect the intended attribute. If not, the dialogue undergoes either improvement or regeneration: improvement refines the previously generated dialogue, while regeneration produces a new dialogue without referencing the previous one. These two strategies are alternated manually. They can help to balance the effectiveness of incremental

refinement and the need to solve the situation when the prior generation is difficult to enhance. The check-and-refine cycle continues until the dialogue successfully reflects the intended attribute. Once a dialogue  $i$  reflecting attribute  $i$  is successfully generated, we perform a coherence check to detect any conflicts\* between *related attributes* and dialogue  $i$ . If the check fails, the intended user attributes will be removed from the dataset. After passing this check, the process moves to the next step.

## 4 Evaluation of IP Capability

### 4.1 Evaluation Framework for IP Ability

Our evaluation framework systematically assesses IP in agent dialogue across three key dimensions:

**Attribute Type Determination.** Given a user history with information on *hidden attribute types*  $T$  (e.g., [age, health, hobby]), an IP-capable model should identify which attribute types are most helpful to the current user question. To quantify this

\*For example, dialogue  $i$  “My grandkids buy me a beautiful dress” conflicts with the *related attributes* {gender: female, age: child}, as a child cannot have grandchildren.

capability, we denote the predicted *related attribute types* as  $T_s$  and introduce *attribute type F1 (ATF)*:

$$ATF = F_1(T_s) = \frac{2 \cdot \text{Precision}(T_s) \cdot \text{Recall}(T_s)}{\text{Precision}(T_s) + \text{Recall}(T_s)}.$$

**Attribute Value Inference.** After identifying the *related attribute types*, models need to predict their corresponding attribute values correctly. Let  $T_s^*$  denote the ground-truth *related attribute types* (e.g., [age, hobby]), and  $A_s^*$  denote the ground-truth *related attributes* (e.g., {age: child, hobby: music}). Each attribute type in  $T_s^*$  corresponds to exactly one attribute value in  $A_s^*$ , so  $|T_s^*| = |A_s^*|$ . We propose the *relative value accuracy (RVA)* score:

$$RVA = \frac{|A_s \cap A_s^*|}{|T_s \cap T_s^*|} = \frac{|A_s \cap A_s^*|/|A_s^*|}{|T_s \cap T_s^*|/|T_s^*|} = \frac{\text{Recall}(A_s)}{\text{Recall}(T_s)}.$$

Among the correctly identified *related attribute types*, the *RVA* measures the proportion of their corresponding attribute values that are accurately predicted.

**Response Generation.** We evaluate response generation using both conventional metrics and LLM-based assessment. Conventional metrics, such as F1-score, offer efficient and deterministic evaluation, while LLM-based assessment enables customized evaluation as well as providing unified scores across different task formats. For conventional metrics, we define *task accuracy* as classification accuracy for **binary-choice** tasks, F1 score for **multiple-choice** tasks, Kendall’s Tau coefficient for **ranking** tasks, and METEOR score for **open-ended** tasks. For LLM-based assessment, we introduce *GPT-4o-Score*, which uses GPT-4o for evaluation. Referencing previous works on LLM-as-a-Judge (Zheng et al., 2023; Cui et al., 2024), we define four key criteria: **conciseness** (0-1 points), **personalization** (0-4 points), **analysis quality** (0-4 points), and **answer accuracy** (0-5 points). For the evaluation prompt, see Appendix C.1.

## 4.2 Reasoning Pathways for IP

Next, we investigate the reasoning process of models on IP tasks. We formalize the five most common reasoning pathways as causal graphs (Pearl, 2009) and design their corresponding Chain-of-Thought (CoT) (Wei et al., 2022) prompts. Each pathway embodies a different hypothesis on how models should process user attributes. Shown in Figure 3, the five reasoning pathways are: (1) **DirectResponse** – the simplest approach, where the model

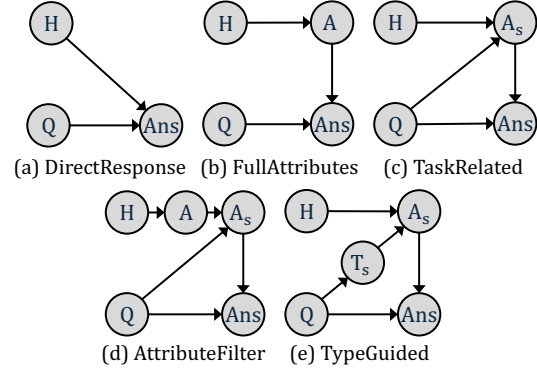


Figure 3: Five reasoning pathways represented as causal graphs.  $H$  represents *user history*,  $Q$  denotes *user question*,  $A$  indicates *hidden attributes*,  $A_s$  refers to *related attributes*, and  $T_s$  is the *related attribute types*.

generates a *response* without explicitly considering the user attributes. (2) **FullAttributes** – the model first predicts all *hidden attributes*  $A$  of the user, then leverage these attributes to generate the *response*. (3) **TaskRelated** – the model directly identifies *related attributes*  $A_s$  before generating the *response*. (4) **AttributeFilter** – the model first predicts *hidden attributes*  $A$ , then extracts *related attributes*  $A_s$ , and finally generates the *response*. (5) **TypeGuided** – the model first infers *related attribute types*  $T_s$ , then predicts specific *related attributes*  $A_s$ , and ultimately provides the *response*.

## 5 Experiments

We begin our experiments with model performance evaluation across three dimensions. Then, we investigate the influence of different reasoning pathways. After that, we analyze the effectiveness of supervised fine-tuning with our training set. Finally, we conduct automatic and human quality evaluation to prove the reliability of our synthetic dataset and its alignment to real-world user conversations.

### 5.1 Setup

**Model.** We evaluate six leading LLMs: GPT-4o (OpenAI, 2023), GPT-o1 mini (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024), Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Meta Llama, 2024), and Qwen2.5-7B-Instruct (Team, 2024).

**Metric.** We use the four metrics in Section 4 for evaluation: *attribute type F1 (ATF)*, *relative value accuracy (RVA)*, *task accuracy*, and *GPT-4o-Score*.

### 5.2 Performance Evaluation

We perform performance evaluations based on our proposed metrics, with more detailed results and



Method	Scenario	Recommendation system			Behavior analysis				Action guide			Average
	Task	Rec	Rank	Fil	Pred	PI	RD	II	Adv	Dec	Conv	
	Answer format	O	R	M	B	M	O	O	O	B	O	
Random		0.00	4.29	11.00	50.00	19.25	0.00	0.00	0.00	50.00	0.00	13.45
GPT-o1 mini		25.13	53.53	66.27	65.20	65.69	27.92	29.85	33.94	73.40	23.63	46.46
GPT-4o		29.43	65.07	64.06	67.40	64.31	29.88	31.22	37.72	76.80	29.89	49.58
Claude-3.5-Sonnet		31.55	61.98	67.43	62.40	67.71	31.13	31.81	36.75	75.00	33.07	49.88
Llama-3.1-70B-Instruct		21.82	42.33	42.80	54.40	42.31	21.98	24.23	26.00	59.60	25.81	36.13
Llama-3.1-8B-Instruct		21.46	42.68	49.25	56.80	38.04	24.85	28.39	33.69	60.80	25.21	38.12
Qwen2.5-7B-Instruct		24.97	52.57	48.61	63.80	50.57	23.11	25.32	32.18	70.80	25.58	41.75
Baseline Avg.		25.73	53.03	56.40	61.67	54.77	26.48	28.47	33.38	69.40	27.20	43.65
SFT-Full		35.15	57.63	69.68	75.80	71.02	47.15	38.96	36.26	83.80	36.29	55.17
SFT-w/o Rec-Fil-Dec		32.70	58.53	61.07	70.60	69.36	47.59	39.32	35.34	80.20	35.70	53.04
SFT-w/o B		35.56	59.14	67.87	37.40	70.16	48.28	39.47	35.29	8.60	35.14	43.69

Table 2: Average *task accuracy* across all reasoning pathways. “O” represents **open-ended**, “R” represents **ranking**, “M” represents **multiple-choice**, and “B” represents **binary-choice**. “Baseline Avg.” stands for the average *task accuracy* of the six non-fine-tuned baselines. “SFT-Full”, “SFT-w/o Rec-Fil-Dec”, and “SFT-w/o B” correspond to Llama-3.1-8B-Instruct fine-tuned on *Full*, *w/o Rec-Fil-Dec*, and *w/o B* training datasets, respectively. For each task, we highlight the **highest score**, the **lowest score**, and the **highest score among non-fine-tuned models**. Note that if the highest overall score is achieved by a non-fine-tuned model, only the blue highlight is used.

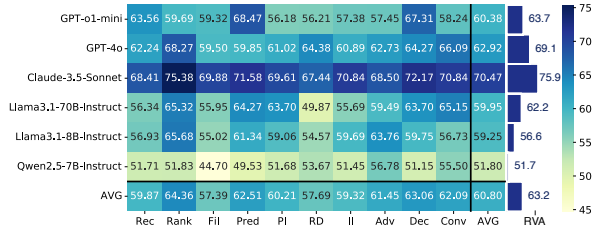


Figure 4: *Attribute type F1 and relative value accuracy*. The heatmap illustrates the *ATF* across models and tasks. The bar chart on the right shows the average *RVA* for each model. The two metrics exhibit a strong positive correlation with Pearson’s correlation of 0.957.

analysis provided in Appendix D.1.

**Attribute Performance.** To assess model capabilities in determining and inferring attributes, we evaluate their average performance across three reasoning pathways: **TaskRelated**, **AttributeFilter**, and **TypeGuided**. These pathways are selected because they all consider extracting the *related attributes* from history explicitly ( $H \rightarrow A_s$ ). Figure 4 presents the *ATF* and *RVA* results, it reveals that: (1) **Strong positive correlation exists between ATF and RVA**: Models with higher *ATF* also achieve higher *RVA*, with Pearson’s correlation coefficient reaching 0.957. This suggests that strengthening either capability naturally possibly enhances the other. (2) **Ranking is the easiest task. Filtering is the most challenging.** (3) **Claude-3.5-Sonnet is the top on both metrics.**

**Task Accuracy.** Table 2 reports the *task accuracy* across models and tasks. We find that: (1) **All mod-**

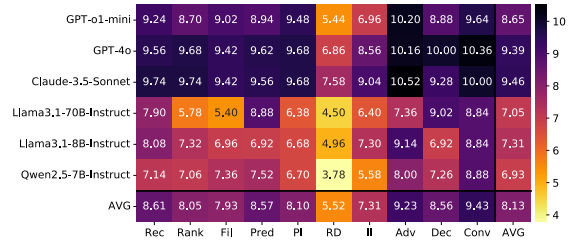


Figure 5: *GPT-4o-Score* across models and tasks, averaged on all reasoning pathways. GPT-4o scores model responses from 0-14 based on criteria in Section 4.1.

**els perform significantly above random guessing**, indicating their fundamental IP capability. (2) **Claude-3.5-Sonnet achieves the highest average task accuracy**, outperforming other models across most tasks. (3) **A correlation emerges between attribute cognition and task performance**: Among the top three models, their ranking in *task accuracy* (Claude-3.5-Sonnet > GPT-4o > GPT-o1 mini) aligns with their ranking in *ATF*, suggesting that stronger attribute recognition contributes to task accuracy in high-performing models.

**GPT-4o-Score.** Shown in Figure 5, from a model perspective, (1) **GPT-4o and Claude-3.5-Sonnet achieve the highest average GPT-4o-Scores**, while (2) **the two Llama models often produce invalid or meaningless responses**, particularly in Llama-3.1-70B-Instruct. From a task perspective, (1) **models generally perform well on convincing but struggle with risk detection.** (2) **Among all scenarios, the difficulty ranking is: behavior analysis > recommendation sys-**

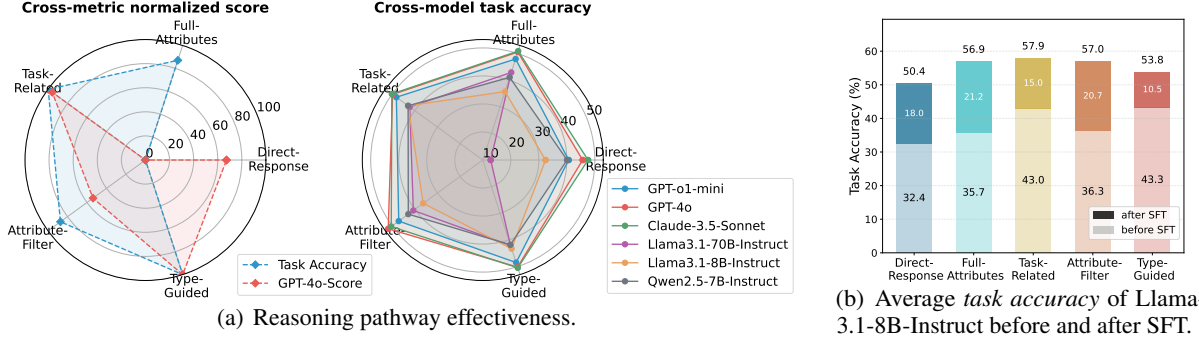


Figure 6: Reasoning pathway evaluation. (a) The left sub-figure compares the normalized average *task accuracy* and *GPT-4o-Score* across five reasoning pathways, averaged on all models. The right sub-figure presents the average *task accuracy* of models under each pathway. **TypeGuided** and **TaskRelated** demonstrate the highest effectiveness. (b) We use **TaskRelated** as the ground-truth reasoning pathway for SFT, considering its efficacy and conciseness.

**tem>action guide.** This aligns with scenario characteristics: **behavior analysis** requires understanding complex psychological factors; **recommendation system** focuses on more concrete matching; the subjective nature of **action guide** leads to conservative scoring by AI judges. Successfully solving these hard scenarios will have substantial model performance gains. Moreover, error analyses on hard scenarios can develop a deeper understanding of how models interpret human behavioral patterns.

### 5.3 Influence of Reasoning Pathways

The left-hand side of Figure 6(a) visualizes the average performance of the five reasoning pathways in *task accuracy* and *GPT-4o-Score*, following min-max normalization for each metric. Combined with the model-specific performance on the right-hand side, we find that: (1) **TypeGuided consistently demonstrates superior performance across both metrics, followed closely by TaskRelated.** This indicates that the extraction of *related attributes* is crucial for effective IP reasoning. (2) **FullAttributes and DirectResponse exhibit high variance between *task accuracy* and *GPT-4o-Score*,** suggesting that certain pathways may perform inconsistently across different evaluation criteria. This variance is understandable given that *task accuracy* evaluates only the final *answer*, whereas *GPT-4o-Score* considers both *analysis* and *answer* quality. Our examination of model outputs reveals that more detailed *analysis* tends to introduce logical inconsistencies and verbosity, which negatively impacts GPT-4o’s “conciseness” and “analysis quality” ratings. As a result, FullAttributes excels in *task accuracy* but underperforms in *GPT-4o-Score*, while DirectResponse demonstrates the reverse trend. (3) **The effectiveness of certain reasoning pathways appears highly**

**dependent on the model’s fundamental capabilities.** This is particularly evident in **DirectResponse**, where weak models struggle significantly. (4) **High-performing models demonstrate less dependency on specific reasoning pathways,** indicating greater robustness in handling implicit personalization tasks. However, their performance with **DirectResponse** remains significantly weaker than other pathways.

### 5.4 SFT on the Training Set

We fine-tune Llama-3.1-8B-Instruct using our training set of 10,790 items. The input consists of *user history*, *user question*, and task instructions, assembled by the prompt of the **TaskRelated** reasoning pathway. The targeted output follows the format of *[related attributes][analysis][answer]*, where the model first identifies relevant attributes, provides reasoning, and then generates the final response. The **TaskRelated** reasoning pathway is selected for two reasons: (1) It achieved top-2 performance across all metrics; (2) It is more direct and less complex than other top-performing pathways (**TypeGuided** and **AttributeFilter**), balancing simplicity with performance.

To further analyze the model adaptability across tasks and answer formats after SFT, we construct three datasets: (1) **Full**: The original training dataset. (2) **w/o Rec-Fil-Dec**: To evaluate the adaptability in unseen tasks, we exclude three tasks (**recommendation**, **filtering**, and **decision**) from **Full**. (3) **w/o B**: To assess the adaptability in unseen answer formats, we remove all **binary choice** tasks: **predicting** and **decision**.

We evaluate the fine-tuned models across all reasoning pathways and report the average *task accuracy* in Table 2: (1) **SFT-Full significantly outperforms the original model**, achieving the highest

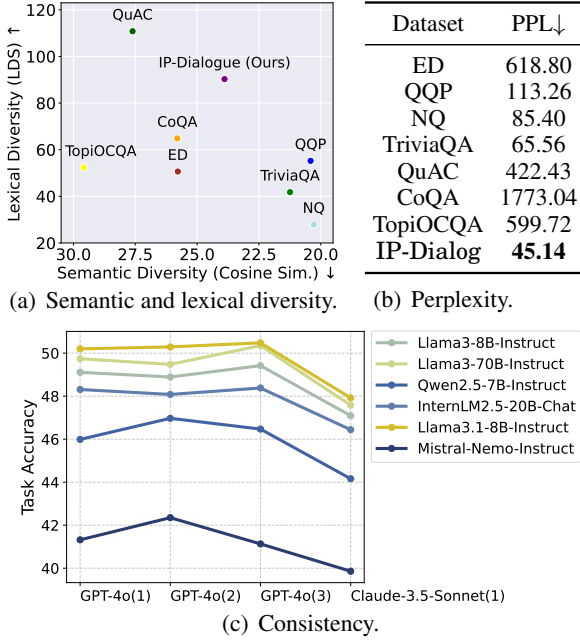


Figure 7: Automatic quality analysis. IP-Dialog shows high semantic and lexical diversity (a) and superior linguistic fluency (b). Additionally, our dataset generation method achieves stable performance assessments across various generation models (c).

scores among all models, proving the effectiveness of SFT on improving IP capability. (2) **SFT-w/o Rec-Fil-Dec generalizes well to unseen tasks**, indicating that the fine-tuned model can adapt well to new tasks with familiar answer formats. (3) **SFT-w/o B exhibits severe performance degradation on binary-choice tasks**, failing in unfamiliar answer formats. Analysis of its *responses* shows that it tends to default to familiar formats from training rather than adopting the required new ones. To address this sensitivity, future training should incorporate more diverse formats. (4) **Training with TaskRelated pathway enhances performance across other reasoning pathways**, demonstrating adaptability in reasoning patterns (Figure 6(b)).

## 5.5 Automatic Quality Evaluation

We conduct three automatic analyses. For more details, see Appendix D.4.

(1) **Diversity**: We use NV-Embed-v2 (Lee et al., 2024) for embeddings of each question, and then calculate the average cosine similarity. A lower cosine similarity indicates greater semantic diversity. We employ MATTR (Covington and McFall, 2010), MTL D (McCarthy, 2005), and HD-D (McCarthy and Jarvis, 2010). We normalize and average these three metrics for a final lexical diversity

Dataset	PPL↓
ED	618.80
QQP	113.26
NQ	85.40
TriviaQA	65.56
QuAC	422.43
CoQA	1773.04
TopiOCQA	599.72
IP-Dialog	<b>45.14</b>

Experiment	Samples	Ann.	Ann. per sample
Attribute Inference	100	2	1
Task Accuracy	50	4	2
Fidelity	200	6	3
Attribute-Dialogue Align.	200	4	2
Attribute-Response Align.	200	4	2

Table 3: Human study setting. “Ann.” means annotators.

score. Shown in Figure 7(a), IP-Dialog achieves leading semantic and lexical diversity.

(2) **Fluency**: We evaluate fluency with the perplexity score (Lai et al., 2022) using Llama-3.1-8B-Instruct. With the lowest PPL score, IP-Dialog exhibits high fluency (Table 7(b)).

(3) **Consistency**: To assess benchmark reliability, we examine whether performance rankings remain stable across different dataset generation models. Figure 7(c) confirms this consistency, as the ranking of *task accuracy* remains robust, validating the reliability of our benchmark.

## 5.6 Human Study

We conduct human studies to evaluate human performance and dataset quality (Appendix D.5). The setting of annotator number is shown in Table 3.

For human performance, in (1) **Attribute Inference Accuracy**, annotators are tasked with inferring attribute types and values from each of the historical dialogues under a predefined set of possible attributes in Table 5. In (2) **Task Accuracy**, annotators answer questions based on a set of *candidate attributes*. For both experiments, **human annotators perform better than or comparably with Llama3-70B-Instruct, but worse than GPT-4o and Claude-3.5-Sonnet**. This reflects our tasks’ high cognitive demands: IP-tasks require advanced reading comprehension, attention to subtle details, and extensive world knowledge, where LLMs have more advantages than humans - our human annotators consistently reported difficulty with the lengthy dialogues and additional background knowledge required for attribute inference. The breakthrough enables the development of reliable LLM-driven personalization services to reduce human efforts.

For quality analysis, (1) **Fidelity**: The human-AI discrimination test for annotators yields an accuracy of 52.2%, indicating that our dataset is nearly indistinguishable from human-generated data. (2) **Attribute-dialogue Alignment**: Human reviewers find that 92.0% of utterances accurately reflect their corresponding ground-truth attributes, demonstrat-



Dataset	Source	Context Type	User Attributes	Implicit Inference
LaMP (Salemi et al., 2024)	Aggregation	Sequence	×	×
NQ (Kwiatkowski et al., 2019)	Crowdsourced	QA	×	×
GlobalOpinionQA (Durmus et al., 2023)	Survey	QA	✓	×
GIE-Bench (Wang et al., 2024b)	Synthetic	QA	✓	×
SynthPAI (Yukhymenko et al., 2024a)	Synthetic	Comment	✓	×
PersonaChat (Zhang et al., 2018)	Crowdsourced	Dialogue	×	×
EmpatheticDialogues (Rashkin et al., 2019)	Crowdsourced	Dialogue	×	×
IN3 (Qian et al., 2024)	Synthetic	Dialogue	×	✓
PersuasionForGood (Wang et al., 2019)	Crowdsourced	Dialogue	✓	×
TopDial (Wang et al., 2023b)	Synthetic	Dialogue	✓	×
<b>IP-Dialog (Ours)</b>	Synthetic	Dialogue	✓	✓

Table 4: Comparison between IP-Dialog and existing datasets. **Context Type**: primary format of the data (QA, dialogue, or preference sequence). **User Attributes**: whether the dataset includes user characteristics (e.g., income level, profession). **Implicit Inference**: whether the dataset requires reasoning from implicit information in context.

ing the high reliability of our dataset. (3) **Attribute-response Alignment**: Annotators assess the consistency between *responses* and *related attributes*, as well as the logical coherence of *analysis*. Among the evaluated samples, 91.9% meet these assessment standards, confirming that our dataset aligns with real user requirements.

## 6 Related Work

**Personalization on Implicit Inference.** Recently, Jin et al. (2024) introduced the concept of implicit personalization (IP), which involves inferring user backgrounds from their queries and tailoring responses accordingly. Current research related to IP is limited. A possible related research direction is user intention understanding (Qu et al., 2018; Cai and Chen, 2020; Kuo and Chen, 2023b; Qian et al., 2024), but they rely on the explicit user answers. Table 4 presents a detailed comparison between existing personalization work and ours.

**Personalization on Explicit Information and Historical Sequences.** While distinct from implicit personalization, other research in personalization offers valuable insights. One line of research focuses on explicit information-based personalization Hovy (2015); Jang et al. (2022); He et al. (2024). Another line of research focuses on analyzing the historical sequences of users to predict future behaviors (Sasaki et al., 2018). Among them, LaMP (Salemi et al., 2024) aggregates 7 tasks for LLM personalization, serving as dataset in many follow-up studies (Zhuang et al., 2024; Liu et al., 2024; Kumar et al., 2024; Tan et al., 2024)

## 7 Conclusion and Discussions

We provide a comprehensive view of implicit personalization. Through an efficient and controllable generation pipeline, we create the IP-Dialog benchmark alongside a training dataset. We develop an evaluation framework featuring four primary metrics and design hypothesized causal graphs to investigate potential reasoning pathways in IP. With extensive experiments, we provide insightful findings and prove our dataset’s reliability.

## 8 Limitations

Though we make our best effort to include as many tasks and user attributes as possible, some of the values are not covered. The limited user attribute design is due to trade-offs between synthetic cost and diversity coverage, as expanding attributes like neurodivergence or intersectional identities demand exponential efforts. Moreover, though our experiments have proved the reliability and fidelity of our datasets, we admit that there could be a potential discrepancy between synthetic dialogues and the real-world user conversations. Due to huge human efforts and time costs to gather such real-world data, we leave this problem to our future work. Finally, we must acknowledge the potential risks associated with the advance of IP technology. IP systems might cause societal stereotypes or biases. To mitigate these risks, we suggest incorporating bias control techniques and restrictions to avoid stereotypes and discrimination.

## 9 Acknowledgment

We would like to thank the anonymous reviewers and the area chair for their valuable feedback, which has significantly improved the quality of this

article. This work was funded by the Shanghai Artificial Intelligence Laboratory.

## References

- Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topicqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.
- Icek Ajzen. 1985. From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior*, pages 11–39. Springer.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. [Open-domain question answering goes conversational via question rewriting](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 520–534. Association for Computational Linguistics.
- Anonymous. n.d. Why do girls sometimes call their female friends girlfriends? why do guys never call their male friends boyfriends? Quora.
- Anthropic. 2024. Introducing the next generation of claude.
- Flávio S Azevedo. 2013. The tailored practice of hobbies and its implication for the design of interest-driven learning environments. *Journal of the Learning Sciences*, 22(3):462–510.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1007–1014.
- D Caroline Blanchard, Guy Griebel, Roger Pobbe, and Robert J Blanchard. 2011. Risk assessment as an evolved threat detection and analysis process. *Neuroscience & Biobehavioral Reviews*, 35(4):991–998.
- Wanling Cai and Li Chen. 2020. [Predicting user intents and satisfaction with dialogue-based conversational recommendations](#). In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020, Genoa, Italy, July 12-18, 2020*, pages 33–42. ACM.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. [Quantifying memorization across neural language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.
- Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024a. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.
- Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu. 2024b. Clear: Can language models really understand causal graphs? *arXiv preprint arXiv:2406.16605*.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentaoh Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [QuAC: Question answering in context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.
- Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucuri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498*.
- Josep-Maria Cots. 1992. Tannen, d. (1991): You just don’t understand. women and men in conversation. *Sintagma: revista de lingüística*; Vol.: 4, 4.
- Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [ULTRA FEEDBACK: boosting language models with scaled AI feedback](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Huy Dao, Yang Deng, Dung D Le, and Lizi Liao. 2024. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 785–795.
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin,

- Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388.
- Lucie Flek. 2020. [Returning the N to NLP: Towards contextually personalized classification models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.
- Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chatrec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.
- Isabel Goddard. What does friendship look like in america?
- Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.
- Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: A factorization-machine based neural network for CTR prediction. In *IJCAI*, pages 1725–1731. ijcai.org.
- Jesse Harte, Wouter Zorndrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1096–1102.
- Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2024. [Cos: Enhancing personalization and mitigating bias with context steering](#). *Preprint*, arXiv:2405.01768.
- Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. [Neural collaborative filtering](#). *CoRR*, abs/1708.05031.
- Dirk Hovy. 2015. [Demographic factors improve classification performance](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.
- Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuseok Lim. 2022. [Call for customized conversation: Customized conversation grounding persona and knowledge](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10803–10812. AAAI Press.
- Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and Athanasios V. Vasilakos. 2013. [Understanding user behavior in online social networks: a survey](#). *IEEE Communications Magazine*, 51(9):144–150.
- Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. [Implicit personalization in language models: A systematic study](#). *CoRR*, abs/2405.14808.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *ACM Sigir Forum*, volume 37, pages 18–28. ACM New York, NY, USA.
- Minbeom Kim, Hwanhee Lee, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2024. [Advisorqa: Towards helpful and harmless advice-seeking question answering with collective intelligence](#). *arXiv preprint arXiv:2404.11826*.
- Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. 2024. [Longlamp: A benchmark for personalized long-form text generation](#). *CoRR*, abs/2407.11016.
- Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems—a survey. *Knowledge-based systems*, 123:154–162.
- Hui-Chi Kuo and Yun-Nung Chen. 2023a. [Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258, Toronto, Canada. Association for Computational Linguistics.
- Hui-Chi Kuo and Yun-Nung Chen. 2023b. [Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 249–258. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti,

- Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, Dublin, Ireland. Association for Computational Linguistics.
- Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoneybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. *DailyDialog: A manually labelled multi-turn dialogue dataset*. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4):2065–2073.
- Jianghao Lin, Yanru Qu, Wei Guo, Xinyi Dai, Ruiming Tang, Yong Yu, and Weinan Zhang. 2023. MAP: A model-agnostic pretraining framework for click-through rate prediction. In *KDD*, pages 1384–1395. ACM.
- Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024. *Llms + persona-plug = personalized llms*. Preprint, arXiv:2409.11901.
- Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, Enhong Chen, et al. 2023. Fairlisa: Fair user modeling with limited sensitive attributes information. *Advances in Neural Information Processing Systems*, 36:41432–41450.
- Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *ACL*, pages 1036–1049. Association for Computational Linguistics.
- Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzhi Xu, Yu Su, and Wenpeng Yin. 2024. *MUFFIN: curating multi-faceted instructions for improving instruction following*. In *The Twelfth International Conference on Learning Representations*, ICLR 2024, Vienna, Austria, May 7–11, 2024. OpenReview.net.
- Philip M McCarthy. 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Ph.D. thesis, The University of Memphis.
- Philip M McCarthy and Scott Jarvis. 2010. MtlD, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- Edwin I Megargee. 1976. The prediction of dangerous behavior. *Correctional Psychologist*, 3(1):3–22.
- Meta Llama. 2024. *Introducing Meta Llama 3: The most capable openly available LLM to date*. Accessed: 2024-09-27.
- Per Egil Mjaavatn, Per Frostad, and Sip Jan Pijl. 2016. Adolescents: Differences in friendship patterns related to gender. *Issues in Educational Research*, 26(1):45–64.
- Masahiro Morita and Yoichi Shinoda. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, organised by Dublin City University, pages 272–281. Springer.
- OpenAI. 2023. *GPT-4 technical report*. *CoRR*, abs/2303.08774.
- OpenAI. 2024. *Introducing openai o1-preview*. Accessed: 2024-11-09.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- Dominic Petrak, Thy Thy Tran, and Iryna Gurevych. 2024. Learning from implicit user feedback, emotions and demographic information in task-oriented and document-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.
- Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2671–2679.
- Henry Prakken. 2006. Formal systems for persuasion dialogue. *Knowl. Eng. Rev.*, 21(2):163–188.
- Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. *Tell me more! towards implicit user intention understanding of language model driven agents*. In *Proceedings of the 62nd Annual Meeting of the Association*



- for Computational Linguistics (Volume 1: Long Papers), pages 1088–1113, Bangkok, Thailand. Association for Computational Linguistics.
- Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. [Analyzing and characterizing user intent in information-seeking conversations](#). In [The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018](#), pages 989–992. ACM.
- Indra Kharisma Raharjana, Daniel Siahaan, and Chastine Fatichah. 2021. [User stories and natural language processing: A systematic literature review](#). [IEEE Access](#), 9:53811–53826.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In [Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers](#), pages 5370–5381. Association for Computational Linguistics.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). [Transactions of the Association for Computational Linguistics](#), 7:249–266.
- Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. [BPR: bayesian personalized ranking from implicit feedback](#). [CoRR](#), abs/1205.2618.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. [Lamp: When large language models meet personalization](#). In [ACL \(1\)](#), pages 7370–7392. Association for Computational Linguistics.
- Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. 2018. [Predicting stances from social media posts using factorization machines](#). In [Proceedings of the 27th International Conference on Computational Linguistics](#), pages 3381–3390, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Lucas Shen. 2022. [LexicalRichness: A small module to compute textual lexical richness](#).
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. [Towards expert-level medical question answering with large language models](#). [CoRR](#), abs/2305.09617.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. [Democratizing large language models via personalized parameter-efficient fine-tuning](#). [CoRR](#), abs/2402.04401.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023a. [Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation](#). In [Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023](#), pages 1132–1143. Association for Computational Linguistics.
- Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023b. [Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation](#). [arXiv preprint arXiv:2310.07397](#).
- Jianling Wang, Haokai Lu, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Ningren Han, Shuchao Bi, Lexi Baugher, Ed H. Chi, and Minmin Chen. 2024a. [Llms for user interest exploration in large-scale recommendation systems](#). In [Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024](#), pages 872–877. ACM.
- Leyan Wang, Yonggang Jin, Tianhao Shen, Tianyu Zheng, Xinrun Du, Chenchen Zhang, Wenhao Huang, Jiaheng Liu, Shi Wang, Ge Zhang, Liuyu Xiang, and Zhaofeng He. 2024b. [Giebench: Towards holistic evaluation of group identity-based empathy for large language models](#). [CoRR](#), abs/2406.14903.
- Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). [arXiv preprint arXiv:1906.06725](#).
- Zhilin Wang and Pablo E. Torres. 2022. [How to be helpful on online support forums?](#) In [Proceedings of the 4th Workshop of Narrative Understanding \(WNU2022\)](#), pages 20–28, Seattle, United States. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In [Advances in Neural Information Processing Systems](#), volume 35, pages 24824–24837. Curran Associates, Inc.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In [Proceedings of the Twelfth Language Resources and Evaluation Conference](#). European Language Resources Association.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. [A survey on large language models for recommendation](#). [World Wide Web \(WWW\)](#), 27(5):60.

- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. [Wizardlm: Empowering large pre-trained language models to follow complex instructions](#). In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.
- Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(1):129–142.
- Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online. Association for Computational Linguistics.
- Ilan Yaniv. 2004. Receiving other people’s advice: Influence and benefit. Organizational behavior and human decision processes, 93(1):1–13.
- Koichiro Yoshino, Yoko Ishikawa, Masahiro Mizukami, Yu Suzuki, Sakriani Sakti, and Satoshi Nakamura. 2018. Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In LREC. European Language Resources Association (ELRA).
- Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024a. A synthetic dataset for personal attribute inference. arXiv preprint arXiv:2406.07217.
- Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin T. Vechev. 2024b. [A synthetic dataset for personal attribute inference](#). In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.
- Donghuo Zeng, Roberto S Legaspi, Yuewen Sun, Xinsuai Dong, Kazushi Ikeda, Peter Spirtes, and Kun Zhang. 2024. Counterfactual reasoning using predicted latent personality dimensions for optimizing persuasion outcome. In International Conference on Persuasive Technology, pages 287–300. Springer.
- Mi Zhang and Neil Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In Proceedings of the 2008 ACM conference on Recommender systems, pages 123–130.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.
- Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. [HYDRA: model factorization framework for black-box LLM personalization](#). CoRR, abs/2406.02888.
- Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. [Toolqa: A dataset for LLM question answering with external tools](#). In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

## List of Appendices

<b>A Design Details</b>	<b>15</b>
A.1 User Attributes . . . . .	15
A.2 Tasks . . . . .	15
<b>B Construction Details</b>	<b>16</b>
B.1 User Question Construction . . . .	16
B.2 User History Construction . . . .	20
<b>C Evaluation Details</b>	<b>22</b>
C.1 Evaluation Standard for GPT-4o-Score . . . . .	22
C.2 Prompts for Five Reasoning Pathways . . . . .	22
<b>D Experiment Details</b>	<b>24</b>
D.1 Performance Evaluation . . . . .	24
D.2 Influence of Different Reasoning Pathways . . . . .	24
D.3 SFT on Trainset . . . . .	24
D.4 Automatic Quality Evaluation . .	25
D.5 Human Study . . . . .	26
<b>E AI Assistants In Research Or Writing</b>	<b>29</b>
<b>F Case Study</b>	<b>29</b>

Attribute Type	Attribute Values
age	child, teenager, young adult, adult, middle-aged, elderly
income level	low, middle, high
profession	student, teacher, sales worker, office clerk, factory worker, software engineer, healthcare worker, driver, farmer, cook, cleaner, journalist, retiree, entrepreneur
openness	high, low
conscientiousness	high, low
extraversion	high, low
agreeableness	high, low
neuroticism	high, low
residence	urban, rural
gender	male, female
health	healthy, minor issue, disabled
hobby	reading, cooking, photography, sports, gaming, movies, travel, music, crafts

Table 5: User attributes and values. These attributes are carefully selected to characterize users and their diverse needs and preferences.

## A Design Details

### A.1 User Attributes

We provide the designed 12 attribute types and values in Table 5.

### A.2 Tasks

We further detail our task design considerations and contributions below.

**Recommendation System.** Recently, there is an increasing focus on leveraging large language models to improve recommendation systems (Wu et al., 2024; Bao et al., 2023; Harte et al., 2023). Most current recommendation systems utilize user historical preference series for personalization (Gao et al., 2023; Christakopoulou et al., 2023; Salemi et al., 2024). While prior LLM research has explored conversational agents in recommendation systems (Liu et al., 2020; Dao et al., 2024), the potential relationships between user dialogues, implicit attributes, and latent interests remain largely unexplored. We propose that leveraging implicit user information presents a promising approach for enhancing recommendation quality. This approach could address several persistent challenges in recommendation systems, including the cold start problem (Lika et al., 2014; Lam et al., 2008), lack of recommendation diversity (Kunaver and Požrl, 2017; Zhang and Hurley, 2008), and the limitation in recognizing the potential needs of users (Wang et al., 2024a). By analyzing implicit user attributes, LLMs can identify potential user needs and suggest relevant items without requiring explicit preferences. Subsequently, by leveraging these LLM-generated elements, the system can expand the recommendation results by discovering similar items in the database, delivering convenient, personalized, and rich recommendations to users.

**Behavior Analysis.** Behavior analysis serves as a fundamental cornerstone for improving user-centric services, such as content recommendations and preference-based customization. While traditional methods in behavior analysis typically rely on extensive user data, LLMs can leverage their intrinsic knowledge about the relationship between user attributes and behavior patterns to generate analytical insights. To comprehensively evaluate this capability, we design four representative tasks that cover different aspects of user behavior understanding. Among them, **intention inference** (Kuo and Chen, 2023a; Qian et al., 2024) has been studied

Task	#Domain	#Subject	#Sample	$ A_s $	Len(H/Q)
Recommendation System					
Recommendation	14	65	100	3.46	177.11/17.43
Ranking	14	65	100	4.68	240.09/30.28
Filtering	14	62	100	3.21	164.53/41.3
Behavior Analysis					
Prediction	15	81	100	4.92	255.58/13.54
Preference Inference	17	77	100	4.02	206.31/24.41
Risk Detection	-	-	100	5.0	256.71/8.0
Intention Inference	12	63	100	4.47	240.19/14.44
Action Guide					
Advice	12	65	100	4.36	223.62/18.26
Decision	12	55	100	4.4	223.7/16.7
Convincing	15	69	100	4.42	243.52/14.3
<b>Average</b>	14	67	100	4.29	223.14/19.87

Table 6: Statistics of IP-Dialog. Each row shows the number of domains and subjects, the number of samples, the average number of related attributes ( $|A_s|$ ), and the average length (in words) of history ( $H$ ) and question ( $Q$ ) for each task.

before. However, previous research relies on interactive dialogue, in which the agent asks the user for more specific detail (Qian et al., 2024; Kuo and Chen, 2023a). Developing a system that automatically infers user intent without explicit questioning would greatly enhance user convenience.

**Action Guide.** **Action guide** aims to transform user intentions into concrete actions (Ajzen, 1985) through three complementary elements: generating practical solutions (**advice**), conducting decision analysis (**decision**), and facilitating behavior change (**convincing**). This scenario integrates informational, analytical, and motivational aspects of guidance to bridge the gap between knowledge acquisition and action implementation. Successfully bridging this gap is critical for personalized LLMs.

## B Construction Details

The statistics of IP-Dialog is shown in Table 6. The pseudo-code of user question generation and user history generation is shown in Algorithm 1 and 2. While our tasks were specifically designed to require user background information for appropriate responses, we acknowledge varying degrees of context-dependency across scenarios. In most cases, historical context significantly impacts the ground truth responses of our dataset. However, a very small subset of examples may exhibit low dependency on historical context, particularly in:

- Binary or multiple-choice questions with limited answer options.
- Questions with strong inherent constraints that naturally narrow potential responses.

- Requests where objective reasoning dominates over personalization needs.

We deliberately included such instances to evaluate whether personalization systems can discern when contextual information is necessary versus when it isn’t relevant. Real-world applications naturally contain questions with these varying personalization requirements, and our dataset reflects this authentic distribution. More generation details are explained below.

### B.1 User Question Construction

For ground truth (GT) answer generation and model evaluation, we limit *related attributes* to no more than 5 to reduce complexity and improve accuracy assessment. During the construction of domains and subjects for user questions, the **risk detection** task stands as an exception, as it consists solely of user attributes without domain and subject distinctions.

The generation process involves multiple specialized prompts (prompts for subject, user question, and GT related attribute and response) presented below. Within these prompt illustrations, the content enclosed in  $\{\}$  varies dynamically during generation based on specific tasks, domains, and contextual parameters. Sample values are shown in  $\{\}$  to aid comprehension. Note that during the generation of user questions, we utilized 3 user questions for each subject and user attribute candidate combination.



---

**Algorithm 1** User Question Construction

---

**Require:**

*domains* ▷ Dict of domains for each task  
*tasks* ▷ List of tasks with name, description and requirements  
*subject\_prompt* ▷ Subject generation prompt  
*user\_question\_prompt* ▷ User question generation prompt  
*GT\_prompt* ▷ Attribute/response generation prompt

**Ensure:** *QA\_items*

```
1: # Generate candidate user attributes
2: candidate_attributes_dataset ← attribute_generator()

3: # User question construction
4: QA_items ← []
5: for each task in tasks do ▷ length of tasks is 10
6:   candidate_attributes_loader ← create_iterator(candidate_attributes_dataset)
7:   for each domain in domains[task.name] do ▷ 10-15 domains
8:     subjects ← GPT4o(subject_prompt(task, domain))
9:     for each subject in subjects do ▷ 10 subjects
10:      for k ← 1 to 3 do
11:        candidate_attributes ← candidate_attributes_loader.next()
12:        user_questions ← GPT4o(user_question_prompt(task, domain, subject,
13:        candidate_attributes))
14:        for each user_question in user_questions do ▷ 3 questions
15:          related_attributes, analysis, answer ← GPT4o(GT_prompt(task, user_question,
16:          candidate_attributes))
17:          QA_items.append((task, domain, subject, user_question, related_attributes,
18:          analysis, answer))
19:        end for
20:      end for
21:    end for
22:  end for
23: end for
24: return QA_items
```

---

---

**Algorithm 2** User History Construction

---

**Require:**

<i>regen_improve_list</i>	▷ Regeneration/improvement strategies
<i>dialog_gen_prompt</i>	▷ Dialogue generation prompt
<i>dialog_improve_prompt</i>	▷ Dialogue improvement prompt
<i>attri_dialog_align_prompt</i>	▷ Alignment examination prompt
<i>dialog_consistency_prompt</i>	▷ Consistency check prompt
<i>related_attributes_dataset</i>	▷ Dataset of related attributes generated from Algorithm 1
<i>QA_items</i>	▷ Question-Answer items from Algorithm 1

**Ensure:** *IP\_dialog\_dataset*

```
1: # Extract unique related_attributes combinations from QA items
2: related_attributes_dataset ← set([QA_item[-2] for QA_item in QA_items])

3: # User history construction
4: item_dialogues ← []
5: for each related_attributes in related_attributes_dataset do
6:   dialogues ← [], dialogue ← ""
7:   for i, related_attribute in enumerate(related_attributes) do    ▷ Generate dialogue per attribute
8:     dialogue ← GPT4o(dialog_gen_prompt(related_attribute, dialogue))
9:     for j ← 1 to 31 do                                          ▷ Try up to 31 times
10:      reflected ← GPT4o(attri_dialog_align_prompt(dialogue, related_attribute))
11:      if reflected or j = 31 then
12:        break
13:      else
14:        if regen_improve_list[j] = "regeneration" then
15:          dialogue ← GPT4o(dialog_gen_prompt(related_attribute, dialogue))
16:        else if regen_improve_list[j] = "improvement" then
17:          dialogue ← GPT4o(dialog_improve_prompt(dialogue, related_attribute))
18:        end if
19:      end if
20:    end for
21:    if not reflected then
22:      discard this related_attributes combination
23:    end if
24:    conflict ← GPT4o(dialog_consistency_prompt(dialogue, related_attributes))
25:    if conflict then
26:      discard this related_attributes combination
27:    end if
28:    dialogues.append(dialogue)
29:  end for
30:  item_dialogues.append((dialogues, related_attributes))
31: end for

32: # Map dialogues to QA items
33: IP_dialog_dataset ← map_dialogues_to_QA_items(item_dialogues, QA_items)
34: return IP_dialog_dataset
```

---

### Subject Generation

Generate 10 subjects on the domain: {Sports}. These subjects are for the task: {decision}. The task requires the agent to {make a decision for a user on whether to perform an action or not.}

The agent will leverage the user's attributes to answer the request related to the subject in a personalized way. The user's attributes include: age, income\_level, profession, openness, conscientiousness, extraversion, agreeableness, neuroticism, residence(urban/rural), gender, health(healthy/minor\_issue/disabled), hobby

-Important Note-

- Make your generated subject as diverse as possible. To cover as much as possible, with the greatest possible differences between categories.
- The subject should be applicable to all user attributes, and the subject should not include any hints about attributes.
- The length of the subject should be less than 5 words.
- Output your generated subjects in the format of:

Output subjects:

- subject 1
- subject 2
- ...

-Example Subject-

Suppose you are asked to generate subjects on the domain: {Education},

Output subjects:

- whether to pursue a graduate degree
- acceptance of university offer
- choice of major

-Your Generated Subject-

Now list your generated subject on the domain: {Sports}

Output subjects:

### User Question Generation

You are a request-generation engine. Your mission is to generate 3 requests on the subject: {Sports}-{join a local sports league}. These requests are for the task: {decision}. The task requires an agent to {make a decision for a user on whether to perform an action or not.}

The agent will leverage the user's attributes to answer the request in a personalized way.

-Important Note-

- You should generate requests in a first-person tone.
- Make your generated requests type as diverse as possible. Avoid generating similar requests.
- Do not mention words such as best fit my unique profile, or any other hints about attributes.
- {Give your answer after the user request. Your user request should not contain any hints about the attributes. It should be hard to make a decision.}

- Output your generated requests in the format of:

Output requests:

- request 1 [Answer: ...]
- request 2 [Answer: ...]
- ...

-Example Requests-

Suppose you are asked to generate request on subject: {Education}-{whether to pursue a graduate degree},

User Attributes: {{'age': 'adult', 'gender': 'female', 'income\_level': 'middle\_income', 'profession': 'cook', 'openness': 'low', 'conscientiousness': 'high', 'extraversion': 'high', 'agreeableness': 'low', 'neuroticism': 'high', 'residence': 'rural', 'health': 'minor\_issue', 'hobby': 'cooking'}}

Output requests:

{Should I pursue a graduate degree in music? [Related Attributes: no music background, high conscientiousness, low openness, rural residence, minor health issue. Answer: No. Reason: Based on your current situation, it may not be the best choice for you.]}

-Your Generated Requests-

Now list your generated requests on subject: {Sports}-{join a local sports league}

User Attributes:

{{'age': 'young\_adult', 'gender': 'female', 'income\_level': 'low\_income', 'profession': 'factory\_worker', 'openness': 'low', 'conscientiousness': 'low', 'extraversion': 'low', 'agreeableness': 'low', 'neuroticism': 'low', 'residence': 'rural', 'health': 'healthy', 'hobby': 'cooking'}}

Output requests:

### Related Attribute and Response Generation

You are a helpful agent for the task: {decision}. The task requires agent to {make a decision for a user on whether to perform an action or not.}

You will leverage the user's attributes to respond to the request in a personalized way. The user request you need to respond to is: {Should I join a local soccer league for fun and socializing?}

–Important Note–

– First, list the user attributes you will use to answer the request in Attributes: [...]. Then, provide the reasoning process for your answer in Analysis: [...]. Finally, present your formatted answer in Answer: [...].

– You should select the user attributes that are most relevant to the request. Use no more than 5 attributes. The attributes should be selected from the given user attribute group.

– Your reasoning should be concise and clear.

– {Your answer should be either yes or no.}

– Output your response in the format below, do not omit the [] in your response:

Output:

– Attributes: [...].

– Analysis: [...].

– Answer: [...].

–Example–

User Attribute Group: {{'age': 'adult', 'gender': 'female', 'income\_level': 'middle\_income', 'profession': 'cook', 'openness': 'low', 'conscientiousness': 'high', 'extraversion': 'high', 'agreeableness': 'low', 'neuroticism': 'high', 'residence': 'rural', 'health': 'minor\_issue', 'hobby': 'cooking'}}

{User: Should I pursue a graduate degree in music?

– Attributes: [profession: cook, openness: low, conscientiousness: high, residence: rural].

– Analysis: [As a cook with low openness, pursuing a graduate degree in music represents a significant departure from the user's current profession and comfort zone. High conscientiousness suggests dedication to current work. Rural residence could limit access to music education opportunities.]

– Answer: [no]}

–Your Response–

User Attribute Group: {{'age': 'young\_adult', 'gender': 'female', 'income\_level': 'low\_income', 'profession': 'factory\_worker', 'openness': 'low', 'conscientiousness': 'low', 'extraversion': 'low', 'agreeableness': 'low', 'neuroticism': 'low', 'residence': 'rural', 'health': 'healthy', 'hobby': 'cooking'}}

User: {Should I join a local soccer league for fun and socializing?}

Output:

## B.2 User History Construction

To prepare attributes for user history generation, we extract and aggregate *related attributes* mentioned in the ground truth *responses* of user questions to form a collection of attribute combinations. Duplicate combinations are consolidated to ensure uniqueness within the set. Subsequently, for each unique *related attributes* combination in this set, we generate corresponding user history dialogues.

For the history generation, we implement an iterative approach consisting of 31 generation-examination iterations per step. An example of our manually designed improvement(i)/regeneration(r) choices is i-i-i-t-i-i-...-i-t-i-t. Generated dialogue in each step that fail to meet our consistency criteria is discarded. As such cases only account for a small portion of our generation results, removal proves more efficient than remediation.

The history dialogue generation process encompasses four prompt types: (1) initial history dialogue generation (and regeneration) for step 0 and step 1+, (2) attribute-alignment examination, (3) iterative improvement described, and (4) consistency verification. The generated history dialogues are paired with user questions sharing the same *related attributes* to construct the final dataset.

After dataset construction, we compute cosine similarity scores between user questions across all samples, constructing a subset where all pairwise similarity scores fall below a threshold of 0.6. Then, we randomly sample 1,000 instances from this filtered subset to form the IP-Dialog benchmark, with the remaining samples comprising the training set.

### Dialogue Generation/Regeneration (Step 0)

Generate a single-round dialogue between a user and AI to implicitly reflect the following user attribute:  
{extraversion: low}

Guidelines:

1. Do not explicitly mention the user attributes.
2. The generated dialogue should implicitly reflect the user's attribute, allowing it to be inferred from the context.
3. Ensure the dialogue remains natural, as if between a user and an AI assistant.

Format the dialogue strictly as follows:

User: [Brief user message that implicitly reflects the given attribute]

AI: [Brief AI response that naturally responds to the user's message without referencing the user's attributes]



### Dialogue Generation/Regeneration (Step 1+)

Based on the following dialogue:  
{User: Can you tell me about some quiet places to visit for a relaxing afternoon?  
AI: Sure! How about a scenic park or a cozy library nearby?}

Continue the dialogue for one more round. The continued single-round dialogue should be between a user and AI to implicitly reflect the following user attribute:  
{agreeableness: low}

Guidelines:

1. Do not explicitly mention the user attributes.
2. The generated dialogue should implicitly reflect the user's attribute, allowing it to be inferred from the context.
3. Ensure the dialogue remains natural, as if between a user and an AI assistant.

Format the dialogue strictly as follows:  
User: [Brief user message that implicitly reflects the given attribute]  
AI: [Brief AI response that naturally responds to the user's message without referencing the user's attributes]

### Dialogue Improvement

Given the following dialogue:  
{User: Parks are usually crowded, and libraries can be too quiet. What else is there?  
AI: You might enjoy a botanical garden or perhaps a quiet café with a nice view.}

Modify this dialogue to make the user message implicitly reflect the following user attribute:  
{agreeableness: low}

Guidelines:

1. Do not explicitly mention the user attribute.
2. The user message should implicitly reflect the given attribute, allowing it to be inferred from the context.
3. Maintain a natural, single-round dialogue between a user and an AI assistant.

Format the modified dialogue strictly as follows:  
User: [Brief user message that implicitly reflects the given attribute]  
AI: [Brief AI response that naturally responds to the user's message without referencing the user's attributes]

### Attribute-Dialogue Alignment Examination

Examine the following dialogue:  
{User: Parks are usually crowded, and libraries can be too quiet. What else is there?  
AI: You might enjoy a botanical garden or perhaps a quiet café with a nice view.}

Can you implicitly infer the user's attribute: {agreeableness: low} from the dialogue? "Implicitly" means the attribute can be reasonably inferred without direct mention, based on the user's language, concerns, or behavior in the dialogue.  
Answer (yes/no): [Your answer here]

### Dialogue Consistency Check

Examine the following dialogue:  
{User: Everywhere is full of people, and libraries are just boring. What's left that's even worth bothering with?  
AI: You might enjoy a botanical garden or perhaps a quiet café with a nice view.}

Is this dialogue consistent with the following user attribute(s)?  
{'extraversion': 'low', 'agreeableness': 'low', 'residence': 'rural', 'health': 'healthy'}

Consistency definition: The dialogue content does not contradict any of the listed user attribute(s).

Answer yes or no only.

Answer (yes/no):

## C Evaluation Details

### C.1 Evaluation Standard for GPT-4o-Score

We define the evaluation standard for GPT-4o-Score with 4 criteria, illustrated by prompt below. As this prompt serves as a formal scoring template, we use parameter names in {} rather than specific examples for a cleaner presentation of the evaluation criteria. Due to evaluation costs, we randomly sample 10 items from each 100-item task for GPT-4o-Score evaluation.

#### Evaluation Standard for GPT-4o-Score

Suppose you are a user with the following attributes: {gt\_related\_attributes}. Your request is: {user\_question}. Now evaluate the following analysis and answers to your request:

Agent Analysis: {predict\_analysis}  
Agent Answer: {predict\_answer}

Evaluation Standard:

Conciseness (0-1 points): Answers should be concise and to the point.

Personalization (0-4 points): Does the analysis and answer specifically address the user's request and attributes?

Analysis Quality (0-4 points): Is the analysis logical, coherent, and well-reasoned? Here's an example of a good analysis: {gt\_analysis}.

Answer Accuracy (0-5 points): Does the answer correctly address the request? Ground Truth Answer: {gt\_answer}.

Total Score: Sum of all categories (0-14 points)

Output your score in the following format:

[Conciseness: score, Personalization: score, Analysis Quality: score, Answer Accuracy: score]  
[Total: score]

For example:

[Conciseness: 1, Personalization: 4, Analysis Quality: 4, Answer Accuracy: 5]  
[Total: 14]

Your Output:

ers (e.g., {task}, {user\_history}, {user\_question}) rather than specific examples in these prompts for a clearer presentation of the differences between the five reasoning pathways. The `attribute_dict` stores the attributes from Table 5 in dictionary format.

#### Reasoning Prompt - DirectResponse

You are a helpful agent for the task: {task}. The task requires the agent to {task\_definition}.

You will leverage my history dialogue to respond to my request in a personalized way. My history dialogue is:

{user\_history}

-Important Note-

- {specific\_note\_for\_task}

- First, provide the reasoning process for your answer in Analysis: [...]. Then, present your formatted answer in Answer: [...].

- Output your response in the format below, do not omit the [] in your response:

Output:

- Analysis: [...]

- {extra\_answer\_format\_for\_task}

My Request: {user\_question}

Output:

#### Reasoning Prompt - FullAttributes

You are a helpful agent for the task: {task}. The task requires the agent to {task\_definition}.

You will leverage my history dialogue and my inferred attributes to respond to my request in a personalized way. My history dialogue is:

{user\_history}

You can infer my attributes from the user attribute group: {attribute\_dict}

-Important Note-

- {specific\_note\_for\_task}

- First, infer my attribute value of ALL the attributes in the user attribute group in Full Attribute Values: [...]. Next, provide the reasoning process for your answer in Analysis: [...]. Finally, present your formatted answer in Answer: [...].

- ALL the attribute values should be selected from the given user attribute group.

- Your reasoning should be concise and clear.

- Output your response in the format below, do not omit the [] in your response:

Output:

- Full Attribute Values: [attribute1: value1, attribute2: value2, ..., attribute12: value12], such as [age: child, income\_level: low\_income, ..., hobby: sports]

- Analysis: [...]

- {extra\_answer\_format\_for\_task}

My Request: {user\_question}

Output:

### C.2 Prompts for Five Reasoning Pathways

In this part, we present the detailed prompts used for each reasoning pathway described in Section 4.2. For all pathways, we provide the model with the user dialogue history  $H$  and current question  $Q$ . We show the 5 designed prompts (**DirectResponse**, **FullAttributes**, **TaskRelated**, **AttributeFilter**, **TypeGuided**) below. Similar to the prompt illustration of evaluation standard for GPT-4o-Score in Appendix C.1, we use placeholder

#### Reasoning Prompt - TaskRelated

You are a helpful agent for the task: {task}. The task requires the agent to {task\_definition}.

You will leverage my history dialogue and my inferred attributes to respond to my request in a personalized way. My history dialogue is: {user\_history}

You can infer my attributes from the user attribute group: {attribute\_dict}

-Important Note-

- {specific\_note\_for\_task}
- First, based on my history dialogue and my request, infer related attributes you will use to answer my request in Attributes: [...]. Then, provide the reasoning process for your answer in Analysis: [...]. Finally, present your formatted answer in Answer: [...].
- You should select the user attributes that are most relevant to the request. Use no more than 5 attributes. The attributes should be selected from the given user attribute group.
- Your reasoning should be concise and clear.
- Output your response in the format below, do not omit the [] in your response:

Output:

- Attributes: [attribute1: value1, attribute2: value2, ...], such as [age: child, profession: student]
- Analysis: [...]
- {extra\_answer\_format\_for\_task}

My Request: {user\_question}

Output:

#### Reasoning Prompt - AttributeFilter

You are a helpful agent for the task: {task}. The task requires the agent to {task\_definition}.

You will leverage my history dialogue and my inferred attributes to respond to my request in a personalized way. My history dialogue is: {user\_history}

You can infer my attributes from the user attribute group: {attribute\_dict}

-Important Note-

- {specific\_note\_for\_task}
- First, infer my attribute value of ALL the attributes in the user attribute group in Full Attribute Values: [...]. Then, extract related attributes you will use to answer my request, and put them in Related Attribute Values: [...]. Next, provide the reasoning process for your answer in Analysis: [...]. Finally, present your formatted answer in Answer: [...].
- You should extract the user attributes that are most relevant to the request. Use no more than 5 attributes. ALL the attribute values should be selected from the given user attribute group.
- Your reasoning should be concise and clear.
- Output your response in the format below, do not omit the [] in your response:

Output:

- Full Attribute Values: [attribute1: value1, attribute2: value2, ..., attribute12: value12], such as [age: child, income\_level: low\_income, ..., hobby: sports]
- Related Attribute Values: [related\_attribute1, related\_attribute2, ...], such as [age: child, profession: student]
- Analysis: [...]
- {extra\_answer\_format\_for\_task}

My Request: {user\_question}

Output:

### Reasoning Prompt - TypeGuided

You are a helpful agent for the task: `{task}`. The task requires the agent to `{task_definition}`. You will leverage my history dialogue and my inferred attributes to respond to my request in a personalized way. My history dialogue is: `{user_history}`. You can infer my attributes from the user attribute group: `{attribute_dict}`.  
-Important Note-  
- `{specific_note_for_task}`  
- First, identify what attribute keys are most related to answer my request in Attribute Key: [...]. Then, based on my history dialogue, infer attribute values of your identified attribute, and write it in Attribute Values: [...]. Next, provide the reasoning process for your answer in Analysis: [...]. Finally, present your formatted answer in Answer: [...].  
- You should select the user attributes that are most relevant to the request. Use no more than 5 attributes. The attributes should be selected from the given user attribute group.  
- Your reasoning should be concise and clear.  
- Output your response in the format below, do not omit the [] in your response:  
Output:  
- Attribute Key: [attribute1, attribute2, ...], such as [age, profession]  
- Attribute Values: [attribute1: value1, attribute2: value2, ...], such as [age: child, profession: student]  
- Analysis: [...]  
- `{extra_answer_format_for_task}`  
My Request: `{user_question}`  
Output:

## D Experiment Details

### D.1 Performance Evaluation

We present more evaluations and findings in this section.

**Attribute Performance.** *ATF* variance on tasks varies more slightly than *ATF* variance on models. GPT-o1 mini ranks the third, which aligns with its documented limitations, as these attribute-related capabilities heavily rely on world knowledge.

**GPT-4o-Score.** Most models maintain an average GPT-4o-Score above 7, demonstrating their basic capability in personalization tasks. The two Llama models often output invalid unmeaningful *responses* with template-like patterns (e.g., "- Analysis: [...] - Answer: [...]"), where meaningful content is replaced with "...". Similar problem have been observed in several prior studies (Chen et al., 2024a,b).

### D.2 Influence of Different Reasoning Pathways

The min-max normalization we used to normalized *task accuracy* and *GPT-4o-Score* to the range [0,1] is

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)},$$

where  $x_i$  represents the score of a specific reasoning pathway, and  $\min(x)$ ,  $\max(x)$  are the minimum and maximum scores among all pathways under the same metric.

We further provide the cross-model comparison on *GPT-4o-Score* in Figure 8. The discoveries are similar with those we report in Section 5.3. Both **FullAttributes** (inferring all possible attributes at the beginning) and **DirectResponse** (generating responses without attribute reasoning) show distinct performance patterns across different models. Models with stronger reasoning and information processing capabilities better adapt to these approaches, with some even achieving superior performance under certain metrics.

### D.3 SFT on Trainset

We train Llama-3.1-8B-Instruct with 4 A100 GPU using LLaMA-Factory<sup>†</sup>. Each of the 3 training processes takes 1-1.5 hours. Table 7 shows the hyper-parameters used in our SFT experiments. We use default values without tuning and report results for each evaluation experiment from a single run.

<sup>†</sup><https://github.com/hiyouga/LLaMA-Factory>



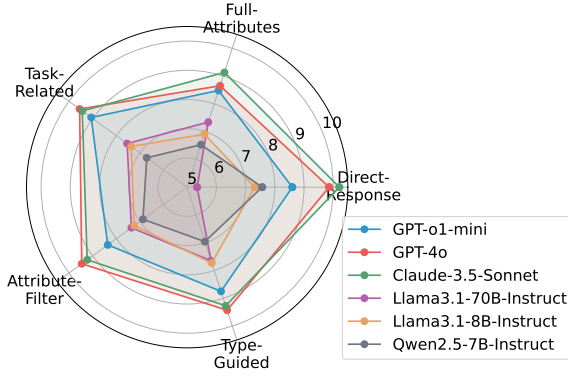


Figure 8: Cross-model *GPT-4o-Score*.

Hyper-parameter	Value
Batch Size	8
Learning Rate	2.0e-5
Number of Training Epochs	3
Warmup Ratio	0.03

Table 7: Hyper-parameters for SFT.

Figure 9 illustrates the *task accuracy* improvements achieved by Llama-3.1-8B-Instruct after SFT across various tasks and CoT prompts. Besides the conclusion in the main paper part, we find that: (1) The performance peak shifts from **TypeGuided** to **TaskRelated** – an expected outcome given the training procedure utilizes the **TaskRelated** reasoning pathway. (2) **TypeGuided**, with its distinctly different reasoning pathway from **TaskRelated**, underperformed compared to prompts that begin with *As* or *A*.

## D.4 Automatic Quality Evaluation

### D.4.1 Baseline

For comparative analysis in diversity and fluency, we select several well-established open-ended QA datasets, including EmpatheticDialogues (ED) (Rashkin et al., 2019), Quora Question Pairs (QQP)<sup>‡</sup>, Natural Questions (NQ)<sup>§</sup>, TriviaQA (Joshi et al., 2017), Question Answering in Context (QuAC) (Choi et al., 2018), Conversational Question Answering (CoQA) (Reddy et al., 2019) and TopiOCQA (Adlakha et al., 2022).

### D.4.2 Diversity

To assess semantic diversity, we utilize NV-Embed-v2 (Lee et al., 2024), a generalist embedding model, to compute embeddings for each question. We then

<sup>‡</sup><https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

<sup>§</sup><https://ai.google.com/research/NaturalQuestions>

calculate the average cosine similarity between all question pairs, where lower mean cosine similarity indicates greater semantic diversity.

For lexical diversity evaluation, we employ three length-insensitive metrics: Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010), Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005), and Hypergeometric Distribution Diversity (HD-D) (McCarthy and Jarvis, 2010). For meaningful comparisons across these metrics with different value ranges, we develop a unified metric called the **Lexical Diversity Score (LDS)**. The LDS formula, defined in Equation (1), normalizes these three metrics to a comparable scale through tangent transformation:

$$\text{LDS} = \left[ \text{mtld} + \tan \left( \text{mattr} \cdot \frac{\pi}{2} \right) + \tan \left( \text{hdd} \cdot \frac{\pi}{2} \right) \right] / 3. \quad (1)$$

The three metrics for lexical diversity evaluation – Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010), Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005), and Hypergeometric Distribution Diversity (HD-D) (McCarthy and Jarvis, 2010) – are computed using the LexicalRichness package (Shen, 2022).

Achieving leading performance in both semantic and lexical diversity in Figure 7(a) demonstrates that IP-Dialog has broad coverage of diverse topics and contexts as well as rich vocabulary.

### D.4.3 Fluency

We evaluate fluency using perplexity scores computed by Llama-3.1-8B-Instruct. Perplexity, defined as the exponentiated average negative log-likelihood of a sequence, serves as a statistical measure of how closely generated texts follow the statistical regularities learned by the given language evaluator. Lai et al. (2022) shows that LLM perplexity correlates with human judgement of fluency, while Wenzek et al. (2020) proposes to use the perplexity of a language model trained on the targeted domain as the quality score. A lower perplexity score indicates that the text follows more natural language patterns.

### D.4.4 Consistency

To evaluate dataset consistency, we randomly sample 1000 items from both the training and test sets (IP-Dialog benchmark) and generate multiple versions of ground truth answers: three from GPT-4o (GPT-4o(1), GPT-4o(2), GPT-4o(3))

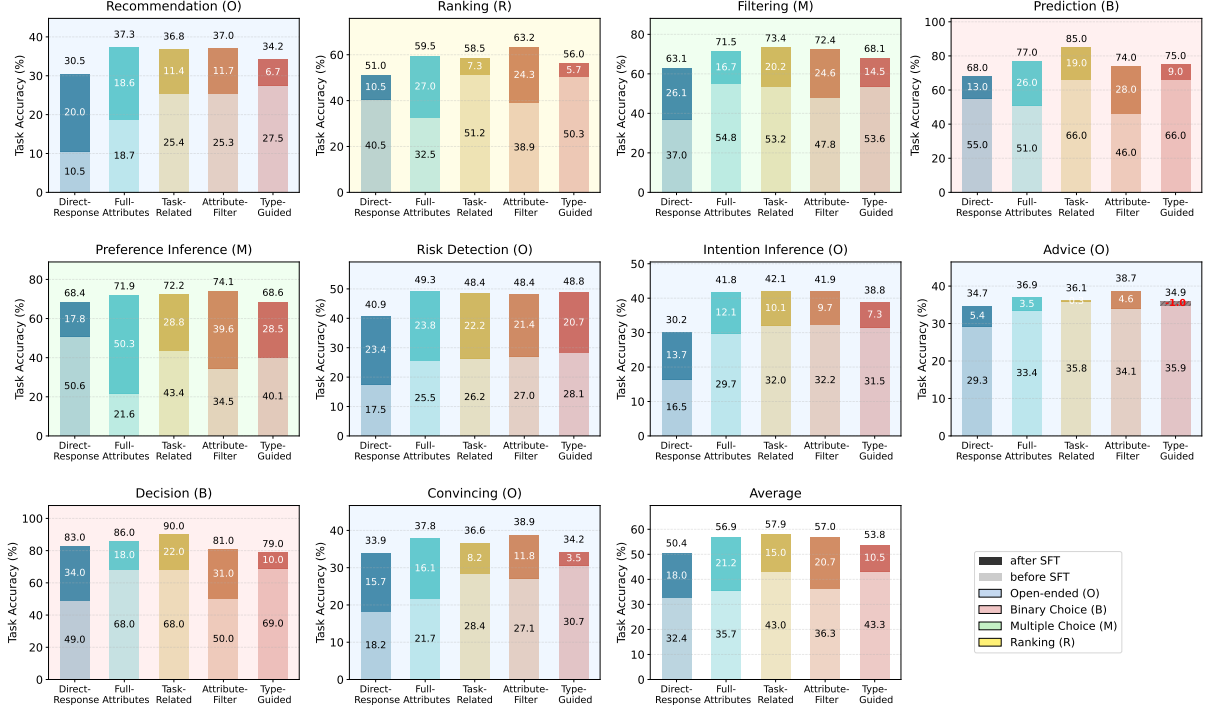


Figure 9: Task accuracy improvement of Llama-3.1-8B-Instruct after SFT.

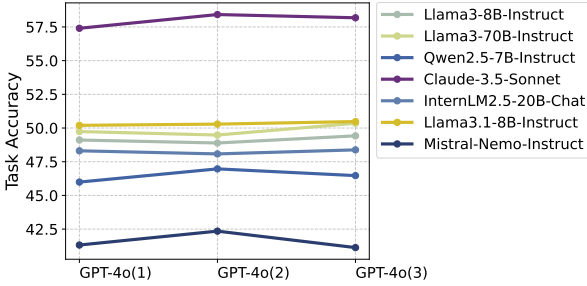


Figure 10: Evaluation consistency check on answer versions of GPT-4o(1), GPT-4o(2), GPT-4o(3).

and one from Claude-3.5-Sonnet (Claude-3.5-Sonnet(1)). We then evaluate six models on these samples: Llama3-8B-Instruct, Llama3-70B-Instruct, Qwen2.5-7B-Instruct, InternLM2.5-20B-Chat, Llama3.1-8B-Instruct, and Mistral-Nemo-Instruct. For each item, the models are provided with hidden user attributes and a user question and are asked to generate a response.

In addition to the consistency analysis results in Figure 7(b), we also provide an analysis of model performances including Claude-3.5-Sonnet on 3 ground truth answer versions in Figure 10, which proves the reliability of our dataset’s ground truth answer.

## D.5 Human Study

In this section, we present detailed information about our human evaluation study, which focused on the average human performance and quality of the generated data. We recruited annotators with diverse backgrounds to conduct the evaluation. All annotators were English-proficient and had at least a bachelor’s degree, ensuring both demographic diversity and academic qualification in our participant pool. The annotators received fair compensation for their work, with all payments funded through our research group. The summarized annotator setting and averaged time used are shown in Table 8.

### D.5.1 Human Performance

**Attribute Inference Accuracy.** To evaluate human performance on attribute inference, we conduct a human study on 100 randomly sampled instances from IP-Dialog. Annotators are tasked with inferring attribute types and values from each of the historical dialogues. The annotation is performed using a predefined set of possible attributes outlined in Table 5. The quality of annotations is assessed using two key metrics:  $Precision(T_f)$  and  $Precision(A_f)$ , where  $T_f$  represents the correctly identified attribute types and  $A_f$  denotes the accurately predicted attribute values. For comparison, we also evaluate the performance of several

Experiment	Samples	Annotator	Ann. per sample	Annotation time per sample	Total annotation time per annotator
Attribute Inference	100	2	1	2.5min	125min
Task Accuracy	50	4	2	3min	75min
Fidelity	200	6	3	15sec	25min
Attribute-Dialogue Align	200	4	2	2min	200min
Attribute-Response Align	200	4	2	2min	200min

Table 8: Human study setting and annotation time.

Method	Human	GPT-4o	Claude-3.5-Sonnet	Llama-3.1-70B	Random
Precision( $T_f$ )	73.2	75.1	<b>80.7</b>	49.5	8.3
Precision( $A_f$ )	62.0	65.6	<b>71.2</b>	38.8	2.0

Table 9: The performances of different methods or models on attribute inference test.

advanced language models (GPT-4o, Claude-3.5-Sonnet, and Llama-3.1-70B) under the same experimental settings. We divided randomly sampled 100 instances into two groups, each independently labeled by one annotator. Recognizing that the annotators might not possess prior knowledge of some attribute types, such as those from the Big Five personality traits, we provided detailed explanations of each attribute to ensure fair evaluation. The final evaluation result is shown in Table 9. In this experiment, GPT-4o and Claude-3.5-Sonnet outperform humans, successfully detecting subtle conversational cues that reflect user attributes. Such capability requires advanced reading comprehension and extensive world knowledge.

**Task Accuracy.** We randomly sampled 50 questions from five distinct close-ended task types: **ranking**, **filtering**, **prediction**, **preference inference**, and **decision**, with 10 instances from each category. These samples were then divided into two groups, with each group independently processed by two reviewers. To assess human performance on these tasks, annotators answered questions based on a set of provided *candidate attributes*. The human annotators achieved an average *task accuracy* of 68.8, comparable to Llama3-70B-Instruct (68.6) but lower than GPT-4o (81.8) and Claude-3.5-Sonnet (76.4).

## D.5.2 Quality Analysis.

**Fidelity** We evaluate whether human annotators could distinguish between AI-generated and human-produced utterances. Our evaluation corpus comprised 100 real dialogues and 100 synthetic dialogues. The real dialogues were sampled from the DailyDialog corpus (Li et al., 2017), which is well-

	Label	Human	AI
Pred.			
Human		201	187
AI		187	112

Table 10: Fidelity analysis: distribution of predictions and true labels in the human-AI utterance classification.

known for its diverse conversational topics and linguistic nuances. The synthetic dialogues were extracted from the user history dialogue of our IP-Dialog dataset. Both sets of dialogues were randomly sampled. To minimize length-related bias, we restricted each dialogue to contain between 25 and 35 tokens, thereby eliminating potential confounding factors that might affect participants’ judgments. The average token count was comparable between the two sets: 29.17 for real dialogues and 29.52 for synthetic dialogues. To maintain objectivity, dialogues were presented to participants in random order, and the source of each dialogue (real or synthetic) was not disclosed. Each dialogue was evaluated by three annotators from a pool of six participants.

As shown in Table 10, participants achieve an accuracy rate of 52.2%, only marginally outperforming random choice. The result indicates that our AI-generated dialogues are nearly indistinguishable from human-generated ones. The inter-annotator agreement, measured by *Fleiss’ Kappa*, was 0.015. This value, being close to zero, indicates minimal consensus among annotators in distinguishing between human and AI-generated content. Such low agreement suggests that our synthetic dialogues achieved a level of naturalness comparable to human-generated ones.

**Attribute-dialogue Alignment.** For the manual evaluation of attribute-dialogue alignment, we randomly sampled 200 instances for review, with each instance assessed by two evaluators. Four evaluators are involved in this experiment. The evaluators assessed whether the dialogue content provided adequate information for attribute inference. They were instructed to flag any instances where attributes could not be reliably inferred and provide brief explanations for these judgments.

92.0% of the utterances are reviewed as accurately reflecting their corresponding ground truth attributes. While our results demonstrate strong overall attribute-dialogue alignment, evaluators identified certain cases where they thought inferring user attributes was too arbitrary. For example:

- In the utterance *"Thanks! My girlfriends keep raving about Notion. Do you know if it has templates for studying or assignment tracking?"*, evaluators questioned whether the use of "girlfriends" sufficiently indicates a female speaker. This hesitation is reasonable, yet sociolinguistic research provides supporting evidence: female speakers statistically use "girlfriends" more frequently than males when referring to female friends (Cots, 1992; Anonymous, n.d.), whereas male speakers typically avoid this term due to its potential romantic connotation, opting instead for "female friends" or simply "friends." Furthermore, social network studies have shown that people typically maintain friendship circles dominated by their own gender (Goddard; Mjauvatn et al., 2016). When someone casually mentions their 'girlfriends' in everyday conversation, it suggests they regularly interact with a female social group. Since people tend to socialize within same-gender circles, this pattern possibly indicates that the speaker is female.
- Similarly, in *"As we review, I can't help but think of this checklist as the script for a blockbuster movie. Every detail needs to be in place for the perfect ending!"*, evaluators questioned whether using movie metaphors indicates film interest. This critical perspective exemplifies thorough evaluation. However, the statement contains multiple film-specific elements: the person naturally uses industry terminology ("script," "blockbuster"), applies film production concepts to everyday

tasks, and references narrative structure ("perfect ending"). When people repeatedly draw metaphors from a specific domain, it typically reflects their familiarity with and interest in that domain. Just as sports enthusiasts often use sports metaphors or musicians use musical analogies, this natural incorporation of film elements suggests some level of engagement with film media.

These examples demonstrate that cases seemingly too ambiguous for attribute inference may contain reasonable linguistic indicators for prediction. In everyday communication, humans also make probabilistic inferences about others based on subtle clues. Our dataset captures this inherent characteristic of human interaction, recognizing both its values and limitations. The evaluators' feedback highlights an important research direction: determining what linguistic patterns constitute sufficient evidence for attribute inference. This is crucial for developing AI systems that understand users naturally and respectfully.

**Attribute-response Alignment.** We measure the degree to which the analysis and responses align with the inferred attributes. With ground truth responses provided, annotators reviewed each GT instance in the dataset for attribute-response alignment. Four annotators are involved to examine 200 samples, with each sample examined by two annotators. They check each user question, its *related attributes*, its ground truth analysis, and answers. The assessment used three key dimensions:

- **Attribute Consistency:** Whether the response properly incorporates and addresses all relevant attributes identified in the analysis phase.
- **Analytical Coherence:** The logical flow between the attribute analysis and the final response.
- **Analysis-Response Consistency:** Whether key insights from the analysis are properly reflected in the final response.

The review process revealed that 91.9% of the evaluated samples demonstrated satisfactory alignment across all assessment criteria. We investigated the sample of errors and found that these instances frequently exhibited inconsistencies between the analysis and the final answer. Specifically, elements emphasized in the analysis are often not given corresponding importance in the final response. This

misalignment suggests potential gaps in the translation of analytical insights into actionable components within the answers.

## **E AI Assistants In Research Or Writing**

This research was conducted with the assistance of AI tools for function documentation lookup during coding and grammar checking during the writing process.

## **F Case Study**

We present five examples from our dataset shown below. Note that the red texts, which provide additional explanations to enhance understanding, do not exist in the original dataset. The blue texts highlight the key content reflecting the *related attribute*.



### Case Study: Recommendation

#### [History]

User: Oh no, I **completely forgot** about my friend's birthday party tonight! Do you think I can grab a last-minute gift on the way there? **[Forgetting important things can be associated with lower conscientiousness.]**

AI: Absolutely! How about a nice gift card or a box of chocolates from a nearby store? You can still make it a great surprise.

User: Great idea! I'm excited to swoop in with something sweet. Plus, it'll be **a chance to chat with everyone**. Do you think I should pick up a card too, or just chocolates? **[People with high extraversion tend to be more willing to chat with others.]**

AI: A card would be a thoughtful touch, and it gives you a chance to write a fun, personal message. It'll definitely add to the celebration vibe!

User: Awesome, I'll **swing by the local shop to stay within budget**. Thanks for helping me make it special without going overboard! **[The user's shopping behavior - being able to shop but showing spending restraint - suggests a moderate income level.]**

AI: Sounds like a perfect plan! Enjoy the party, and I'm sure your friend will appreciate the thoughtful gesture.

User: I might have to **dodge some traffic getting to a shop**, but I'm glad there are plenty of options nearby. **[The presence of traffic congestion in this area suggests an urban rather than rural location.]**

AI: With so many stores around, you're bound to find everything you need quickly. Safe travels, and enjoy the party!

#### [Task]

Recommendation: recommend products based on the needs of a user's specific attribute.

#### [Question]

I'm thinking about getting a car that's suitable for city life. What would you suggest?

#### [Related Attributes]

Conscientiousness: low, extraversion: high, income level: middle income, residence: urban.

#### [Analysis]

Given the user's income level, **practical and budget-friendly** options are recommended. An urban residence suggests a need for a **compact, fuel-efficient** car that's **easy to maneuver and park in city environments**. High extraversion might indicate an interest in a car **with a stylish design and features** that accommodate an active lifestyle. Low conscientiousness suggests a preference for a car that **requires low maintenance**.

#### [Answer]

Honda Fit, Toyota Yaris, Ford Fiesta. **[These are compact, fuel-efficient, stylish, and low-maintenance car models that are easy to park and suit an active, urban lifestyle on a budget.]**

Figure 11: Case study: Recommendation.

## Case Study: Ranking

### [History]

User: I found this perfect spot with **amazing lighting** yesterday that I can't wait to visit again soon. Any tips on **tracking the best times for natural light** next time I go there? **[Focusing on ideal lighting conditions is a characteristic of photographers.]**

AI: You might want to try using a sun-tracking app to check the position of the sun throughout the day. It can help find the ideal time for the kind of natural lighting you're looking for.

User: Are there any **free** apps you recommend for sun tracking, so I can plan **without spending extra**? **[The emphasis on finding free apps to avoid additional costs suggests a need to minimize expenses, implying a lower income level.]**

AI: Definitely! There are several free apps available, like Sun Surveyor Lite and SunCalc, which should help you track the sun's path without any cost.

User: I might just **stick** to visiting **at the same time** as yesterday. It worked pretty well. **[Preferring to return at the same time instead of experimenting with new options indicates a comfort with routine and familiarity, which suggests low openness.]**

AI: That sounds like a solid plan! Sometimes sticking with what works is the best approach. If you need any other tips, feel free to ask!

User: Maybe I'll capture some photos to show to our team during our next meeting. Visuals could really **make our pitch more compelling** and highlight the presentation with all that natural light. **[The emphasis on using visuals to "make our pitch more compelling" suggests the user may be involved in sales or business development roles where persuasive presentations are common.]**

AI: That sounds like a great idea! Sharing those photos can really help capture everyone's attention and set a positive mood. If you have any other thoughts or need assistance, let me know!

User: It's amazing how **urban landscapes** can create stunning backdrops too, especially with the right lighting. Any tips on capturing the uniqueness of **city architecture** in photos? **[The specific knowledge of urban lighting and architectural details suggests they likely reside in a city environment, where they can regularly observe these elements firsthand.]**

AI: Absolutely! Try focusing on unique angles or perspectives of the buildings, and use reflections in windows or puddles to add an interesting dimension. Playing with different times of day can also help capture the essence of urban life.

### [Task]

Ranking: rank a series of given items according to the user's potential level of interest.

### [Question]

I'm looking to prioritize the following economic topics in terms of how interesting they might be for me:

A. inflation impacts on urban living B. trends in tech startups C. renewable energy investments D. financial literacy programs

### [Related Attributes]

Hobby: photograph, income level: low income, openness: low, profession: sales worker, residence: urban.

### [Analysis]

Given the user's low income level and urban residence, topics related to inflation impacts on urban living and financial literacy programs might be of higher interest as **they could directly affect and benefit their personal situation**. The user's profession as a sales worker might **create some interest in trends in tech startups**, as it can influence market dynamics and job opportunities. The user's low openness suggests a preference for familiar or immediately relevant topics, which would prioritize practical over abstract interests. Lastly, renewable energy investments, while broadly significant, might **not be as immediately relevant or interesting to someone with this profile and hobby in photography**.

### [Answer]

A, D, B, C

Figure 12: Case study: Ranking.

### Case Study: Filtering

#### [History]

User: Hey AI, I just got back from [hosting a dinner party](#), and it was [so much fun](#)! We played games, exchanged stories, and I think everyone [had a great time](#). Any suggestions for what we could do next week to [keep the energy up](#)? [The speaker's enthusiasm for socializing, hosting, engaging in interactive activities, and planning future gatherings reflects high extraversion, as they derive energy from social interactions and enjoy lively group activities.]

AI: That sounds amazing! How about organizing a themed karaoke night? You could even set up a small stage area to really amp up the excitement.

User: That sounds like a blast! I already have a speaker and a laptop, so I think I can make it work without [breaking the bank](#). Any tips on how to make it feel special [without spending too much](#)? [The user's priority on cost-effective solutions over spending on extras suggests a low income level.]

AI: Absolutely! You could create a fun playlist with everyone's favorite songs or decorate the space with some DIY lights and props for a festive touch. Maybe even make a karaoke contest out of it with some homemade certificates as prizes!

User: Great ideas! I can check with my neighbor about [borrowing some hay bales](#) to use for seating, and maybe we can have it outside [under the stars](#). It's [pretty quiet](#) around here, so we shouldn't bother anyone. [The user's mention of borrowing hay bales for seating, hosting an outdoor event under the stars, and describing the area as "pretty quiet" indicates a rural residence, where open space and agricultural resources are more accessible.]

AI: That sounds charming! An outdoor setup will make it memorable, and you could even hang string lights between trees for a cozy atmosphere. Don't forget to have some warm drinks ready if it gets chilly!

#### [Task]

Filtering: filter a given list of items based on user attributes and preferences, retaining suitable content and removing irrelevant items.

#### [Question]

Help me filter these transportation options for tourists. Which ones should I consider?

A. Private yacht rental B. Public subway system C. Guided city walking tour D. Shared ride-hailing service E. Camper van rental

#### [Related Attributes]

Extraversion: high, income level: low income, residence: rural.

#### [Analysis]

The user's low income level suggests [cost-effective](#) transportation options are preferable. High extraversion indicates they may [enjoy social environments](#). Their rural residence might mean they appreciate opportunities to [immerse in more urban or communal settings](#). The public subway system and shared ride-hailing service both align with these considerations, offering affordable and potentially social travel experiences.

#### [Answer]

B, D

Figure 13: Case study: Filtering.

## Case Study: Prediction

### [History]

User: I finished organizing my schedule for the week, making sure each task has a specific time slot. Can you remind me about my morning meeting tomorrow, just in case something changes? [The proactive scheduling, attention to task timing, and desire for reminders reflect high conscientiousness, as they demonstrate thorough planning and a strong sense of responsibility.]

AI: Of course, I'll remind you about your meeting and keep an eye out for any updates. Let me know if there's anything else you need help with!

User: I'll keep to myself for now. [The choice to stay reserved and avoid social interaction indicates low extraversion.]

AI: Alright, feel free to reach out when you want.

User: Actually, could you also check if any updates are needed for the itinerary of my business trip next week? I'm flying first class and want to make sure everything is in order. [The ability to afford first-class airfare for a business trip suggests a high-income level.]

AI: Certainly, I'll verify all the details of your trip itinerary and ensure that everything is up to date and smoothly arranged for you.

User: On second thought, I'd love some recommendations for things to do and places to explore around my destination. I have some free time during my trip and I'm curious to see what I might discover! [The interest in exploring new activities and discovering unfamiliar places during the trip indicates high openness.]

AI: Absolutely! I'll put together a list of suggestions for unique experiences and must-see spots at your destination. Let me know if you want to focus on anything specific!

User: I'm looking forward to some downtime and hoping to explore some serene spots like the ones I'm used to back home. Anything like that available? [The specific mention of finding serenity "like home" during downtime may indicate rural residence, as rural dwellers often seek familiar natural tranquility when visiting cities - a pattern less common among urban residents.]

AI: Certainly, I'll find some serene nature trails and charming local farms at your destination for you to explore. You'll have a peaceful and rewarding experience for sure!

### [Task]

Prediction: predict whether a user will take a specific action or make a particular decision.

### [Question]

Will I opt for a carpool arrangement to reduce my travel expenses?

### [Related Attributes]

Conscientiousness: high, extraversion: low, income level: high income, openness: high, residence: rural.

### [Analysis]

While high income may reduce the need to cut travel costs, a high level of openness suggests you are open to new experiences like carpooling. High conscientiousness may drive you towards efficient and environmentally friendly choices. However, low extraversion might make you less inclined to share rides with others. Living in a rural area may also influence the practicality and availability of carpooling.

### [Answer]

No.

Figure 14: Case study: Prediction.

### Case Study: Convincing

#### [History]

User: I just finished organizing my tasks for the week. Could you remind me to review my plan every morning at 8 AM? [The structured task organization and desire for daily 8 AM reviews indicates high conscientiousness through the disciplined planning habits of the user.]

AI: Sure, I'll set a reminder for you every morning at 8 AM to review your weekly plan.

User: Thanks! I was thinking of adding "plan my next trip" to the schedule. Do you think that's a good idea? The interest of the user in planing the next trip implies that travel is an activity they enjoy and prioritize.

AI: That sounds like a great addition! It could be exciting to explore new destinations as part of your weekly routine.

User: I'm actually a bit worried that I might forget something important during my trip planning. Do you think I should create a checklist to make sure I cover everything? [The expression of worry about forgetting trip details and immediate request for organizational tools like checklists suggests heightened anxiety about potential oversights - a characteristic often associated with high neuroticism.]

AI: Creating a checklist is always a smart move to ensure nothing gets overlooked. I can help you set one up if you'd like!

#### [Task]

Convincing: convince a user to do something.

#### [Question]

Encourage me to avoid single-use plastics during my travels.

#### [Related Attributes]

conscientiousness: high, hobby: travel, neuroticism: high.

#### [Analysis]

The user has high conscientiousness, suggesting they are organized and responsible, which aligns with the goal of reducing plastic waste. Their interest in travel is relevant as it directly relates to the context of avoiding single-use plastics while traveling. Additionally, their high neuroticism might influence their sensitivity to environmental concerns or anxieties about contributing to pollution, which could be addressed by taking positive actions like reducing plastic use.

#### [Answer]

As someone who is conscientious and passionate about travel, you have a unique opportunity to make a positive impact on the environments you explore. By avoiding single-use plastics, you can help preserve the natural beauty of your travel destinations for future visits and generations. Plus, taking this eco-friendly approach could ease any environmental concerns you may have, allowing you to enjoy your travels with peace of mind, knowing you're contributing to a healthier planet.

Figure 15: Case study: Convincing.