Is He Extroverted?
Identifying Missing Relevant Personas for Faithful User Simulation

Weiwen Su[1,3], Yuhan Zhou[*1], Zihan Wang[*1], Naoki Yoshinaga[2,3], Masashi Toyoda[2,3]
[1]The University of Tokyo, [2]Institute of Industrial Science, The University of Tokyo
[3]Institute for Digital Observatory, The University of Tokyo
{su-w, yzhou, zwang, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract

Existing user simulation approaches focus on generating user-like responses in dialogue. They often assume that the provided persona is sufficient for producing such responses, without verifying whether critical personas are supplied. This raises concerns about the validity of simulation results. To address this issue, we study the task of identifying persona dimensions (*e.g.*, "*whether the user is price-sensitive*") that are relevant but missing in simulating a user's reply for a given dialogue context. We introduce PICQ-drama (constructed from TVShowGuess), a benchmark of context-aware choice questions, annotated with missing persona dimensions whose absence leads to ambiguous user choices. We further design diverse evaluation criteria for missing persona identification. Benchmarking leading LLMs on our PICQ-drama dataset demonstrates the feasibility of this task. Evaluation across diverse criteria, along with further analyses, reveals cognitive differences between LLMs and humans and highlights the distinct roles of different persona categories in shaping responses. The dataset is available at:

https://github.com/NioHww/PICQ/

1 Introduction

User simulation aims to model the behavior of a target user in a hypothetical situation and is commonly studied to predict the user's responses given a dialogue context and additional data to characterize the user. Recent large language models (LLMs) have greatly expanded its potential, supporting applications such as non-player characters in games (Park et al., 2023), character-based response generation (Shao et al., 2023; Wang et al., 2024; Tu et al., 2024), and opinion dissemination (Gao et al., 2023). These simulations often rely on rich personas or interaction history, either manually prepared or derived from external sources (*e.g.*, Wikipedia), without considering practical simulation situations.

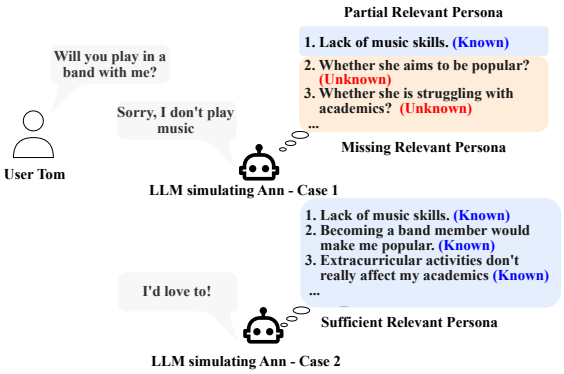

Figure 1: User simulation with sufficient versus partial relevant personas can lead to opposite answers, even when the simulation itself is accurate.

Existing user simulation studies attempt to accumulate as comprehensive personas as possible in advance, using external sources such as biographies (Shao et al., 2023; Wang et al., 2025) or interviews (Ng et al., 2024; Park et al., 2024), to support simulation across a wide range of situations. However, simulations in individual situations are influenced by diverse, situation-specific personas that cannot be fully captured by generic biographies or situation-agnostic interviews. As a result, user simulation based on situation-agnostic comprehensive personas suffers from missing relevant personas (Figure 1), leading to unfaithful user simulation. Moreover, since not all personas are relevant to a given situation, using comprehensive personas for a specific situation can be unnecessarily costly.

In this study, to enable faithful user simulation based on relevant personas, we assume a basic persona (including age, gender, and the interlocutor relationship) and focus on identifying missing personas that are relevant in a specific simulation context. By solving this task using LLMs, we ask the following research questions:

[*]Equal contribution.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 131–149

RQ1: How well can leading LLMs identify missing relevant personas from context?

RQ2: What cognitive patterns emerge in their performance, especially compared to humans?

RQ3: Can effective instruction strategies enhance model performance on this task?

To answer these research questions in a controlled and meaningful setting, we construct a new benchmark dataset, PICQ-drama, based on character-rich drama scripts in the TVShowGuess dataset (Sang et al., 2022). We focus on user responses to persona-influenced choice questions (PICQs), where respondents select from a constrained set of options (*e.g.*, "Would you marry me?"); answers to PICQs naturally reveal user preferences shaped by personas. Empirically, PICQs constitute the majority of persona-influenced questions (§ 3) in the TVShowGuess dataset. Our dataset pairs context-aware PICQs with annotated missing relevant persona dimensions. Annotation proceeds in two stages: LLM pre-screening combined with manual verification identifies PICQs from dialogues; annotators determine and describe the missing persona dimensions that influence each PICQ choice. In addition, we propose a multi-faceted evaluation scheme with three metrics, assessing influence on user choices, the difficulty of acquisition (inaccessibility), and alignment to human-annotations. We also design a multi-task instruction strategy designed for this task.

In our experiments, we benchmark leading LLMs such as GPT-4.1 (OpenAI, 2024), Qwen-3 (QwenTeam, 2025), and Llama-3.1 (LlamaTeam, 2024). The results confirm the feasibility of applying LLMs to this task and validate the effectiveness of our instruction strategy. We evaluate the models from various aspects of influence, inaccessibility, and fidelity with human annotations. We investigate the influence of model scales on the performance, the influence of our instruction strategies, and the cognitive patterns of the models.

Our contributions lie in (1) We formulate a new, query-focused task for identifying missing relevant personas to ensure persona sufficiency; (2) We create the PICQ-drama benchmark with PICQs annotated for missing personas and evaluation metrics; (3) We design a multi-task instruction strategy to enhance the influence of identified personas; and (4) We conduct evaluation and analysis to reveal cognitive differences among LLMs and humans.

2 Related Work

In this section, we first review the literature on user simulation to position our task. We then introduce existing persona-augmented dialogue datasets and clarify how our dataset relates to and differs overall.

2.1 User Simulation

Recent studies using LLMs to simulate human responses can be categorized by persona granularity: demographic, biography, and individualized personas (Chen et al., 2024).

User simulation with demographic persona captures behavior patterns of certain groups (*e.g.*, "a 25-year-old white woman") rather than specific individuals (Deshpande et al., 2023; Kong et al., 2024). While this setting does not aim to simulate a unique individual, it does not eliminate the issue of persona insufficiency. Specifying only demographic attributes is often insufficient to determine a unique response, and the appropriate output in such cases may be a distribution of plausible behaviors rather than a single prediction. Therefore, we focus on individual simulations, where the goal is to approximate a specific person's response, making persona sufficiency a critical concern.

Individual simulation with biography mimics fictional characters or celebrities. This type of target often provides abundant persona data for simulation, such as scripts (Tu et al., 2024; Chen et al., 2023), summarized personas (*e.g.*, from Wikipedia) (Shao et al., 2023), or even parametric knowledge in LLMs (Lu et al., 2024), enabling rich persona input. However, abundant personas do not necessarily imply that the personas relevant to a specific context are available. It remains unclear whether models can effectively utilize the most relevant personas from the available information or recognize when crucial personas are not present.

Individual simulation without biography focuses on real-world individuals for applications such as personalized services. Due to privacy constraints and limited access, the persona information available in advance is often sparse. Prior work has explored various methods to collect persona information (Ng et al., 2024; Park et al., 2024; Yamashita et al., 2023), aiming to gather rich persona of an individual via interviews, pre-specified questions, or questionnaires (*e.g.*, MBTI). However, such methods do not guarantee persona sufficiency in specific contexts. Instead, focusing on query-relevant personas for each situation provides a more practical

approach to ensure the persona sufficiency. Identifying relevant personas for each query is thus crucial for improving simulation faithfulness.

In summary, our work focuses on the ill-posed problem in user simulation caused by insufficient relevant personas. Instead of accumulating more persona data, we ask: "Which persona dimensions are necessary for a given simulation scenario or query?" We formalize this as the task of identifying missing relevant personas in a query-focused context and provide the first benchmark and in-depth analysis for it. By emphasizing minimal yet sufficient persona per query, it follows the "less is more" principle, paving the way toward more well-posed, efficient, and faithful simulations.

2.2 Persona-Augmented Dialogue Dataset

Existing dialogue datasets with personas, such as PersonaChat (Zhang et al., 2018), Multi-Session Chat (Xu et al., 2022), and CharacterEval (Tu et al., 2024), can indeed be used to simulate responses personalized to the target personas. However, these datasets do not provide annotations indicating what personas influence each individual response, making it difficult to study persona sufficiency or to identify missing relevant personas.

In contrast, our PICQ-drama dataset explicitly annotates the missing relevant persona dimensions for each response or decision, enabling controlled evaluation of query-specific persona sufficiency. By providing this fine-grained mapping between queries and the personas that shape the responses to them, our dataset allows models to not only generate user-like responses but also to identify which persona dimension is still missing, enhancing the fidelity and interpretability of user simulation.

3 Query-Focused User Simulation

A key challenge in studying persona sufficiency is that, in general dialogue, the influence of persona on an utterance is often implicit and difficult to isolate. To make this influence explicit and analyzable, we focus on user responses to questions rather than questions themselves or phatic expressions (*e.g.*, greetings, farewells). Answers to questions often reveal information that directly affects the questioner's subsequent decisions.

To understand the types of questions that naturally arise in dialogue, we conducted a manual analysis of 400 randomly sampled sentences ending with a question mark from the TVShowGuess dataset (Sang et al., 2022). We categorized them into three groups: fact-seeking questions (*e.g.*, "What is the definition of quantum?" or "Where are you from?"), persona-influenced questions, and non-questions. Fact-seeking questions accounted for 61.1%, persona-influenced questions for 25.6%, and the remainder were not genuine questions. Although fact-seeking questions constitute the majority, their answers are primarily external facts or personal facts, making them less suitable for studying the influence of persona. We further examined the persona-influenced questions and found that the majority (approximately 75%) are choice-based, where the respondent selects from a small set of alternatives, while the rest are open-ended. This observation suggests that choice-based questions are the dominant form of persona-influenced questions in daily dialogue.

Motivated by this data-driven observation, we focus on these persona-influenced choice questions (PICQs), where answers reflect the targets' decisions or opinions shaped by their personas (*e.g.*, "Would you form a band with me?"). Compared to open-ended questions, PICQs also provide a constrained structure that enables controlled and comparable simulation.

In addition to PICQ, we consider including dialogue context preceding each question as the input query. This is because certain questions may appear simple in form (*e.g.*, "Would you stay here with me?") but derive their significance from complex preceding situations (*e.g.*, "they are outside late at night in the rain"). Without context, it would be difficult to accurately interpret the choice being made or simulate a meaningful response. Moreover, some questions are not self-contained (*e.g.*, "Would you do that with me?") and cannot be interpreted or answered without the surrounding dialogue.

Following this logic, we provide a basic persona description for the responding character, including gender, age, and their relationship to the questioner (*e.g.*, "co-worker" and "stranger"). These attributes are chosen because they are broadly applicable across diverse questions, commonly adopted in persona-augmented datasets and character simulations (Zhang et al., 2018), and serve as stable anchors for inferring more specific persona dimensions relevant to the choice. Here, a persona dimension is defined as a trait axis whose specific value is currently unknown (*e.g.*, *"whether s/he is shy"*). For brevity, we use persona to refer to the persona dimension in the remainder of the paper.

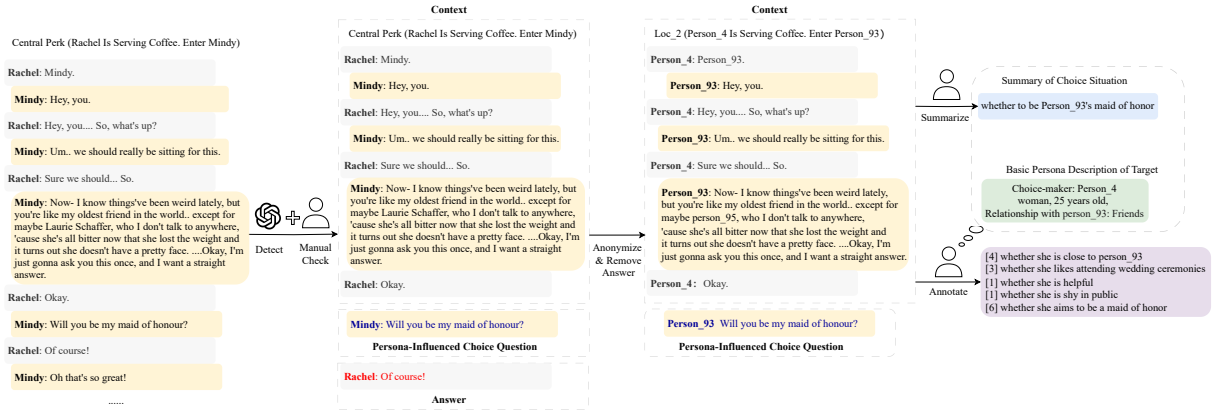

Figure 2: Overview of our approach to acquiring PICQs and annotating the missing relevant personas.

Therefore, we define the task of identifying missing relevant personas as follows:

Input: A dialogue context ($\mathbf{C}$), a PICQ ($\mathbf{Q}$), and a basic persona description ($\mathbf{P}$).

Output: A set of missing relevant personas ($\mathbf{P_{unk}}$) that are likely to influence the answer ($\mathbf{A}$) to the question as the output.

4 Data Collection and Annotation

In this section, we describe our dataset derived from the TVshowGuess dataset (Sang et al., 2022), comprising dialogue contexts, persona-influenced choice question (PICQ), a basic persona description, human-identified missing relevant personas, and answers. It is constructed from the source dialogues in two steps: i) identifying PICQ instances, and ii) annotating the missing relevant personas likely to influence responses, as shown in Figure 2.

4.1 Source Dialogue Dataset

We first select a source dialogue dataset that meets two criteria: i) a realistic setting to support user behavior simulation, and ii) sufficient persona descriptions to help identify missing relevant personas and assess their impact on the simulation task.

Based on the above two criteria, we selected the TVshowGuess dataset (Sang et al., 2022), which contains English scripts from five popular TV sitcoms. We chose three series, including *Friends*, *Frasier*, and *The Office*, balancing topical diversity with annotation feasibility. These series cover themes like friendship, romance, family, and career, aligning well with our focus on everyday choice-making. We use the first three seasons of them as the source dialogues.

4.2 Discovering PICQs and Answers

Our first stage of annotation is to identify PICQs and their answers from the source dialogues. The definition of PICQs is discussed in § 3. A corresponding answer to a PICQ is defined as the immediate next utterance following PICQ that clearly chooses one of the alternatives implied or listed by the question. Restricting the answer to the next utterance ensures that it reflects the respondent's initial persona-based intent and avoids incorporating later choices that may result from discussions.

To reduce annotation cost, we first prompt GPT-4.1 to detect potential PICQs and their answers. Three human annotators (the first, second, and third authors) then verify whether each candidate pair satisfies our task criteria. Refer to Appendix A.4 for the prompts and Appendix A.3 for the annotation guidelines. Each annotator reviews two-thirds of the data to ensure overlap and allow measurement of inter-annotator agreement. The average Cohen's κ between annotator pairs is 0.740, indicating substantial agreement. Disagreements are then resolved by discussion to ensure data quality.

4.3 Annotating Missing Relevant Personas

Our second stage of annotation is to identify missing relevant personas for each PICQ, given its context, and the respondent's basic persona (§ 3). We first define what missing personas can be annotated, and then introduce our annotation process.

Based on the persona definitions from previous work (Chuang et al., 2024; Yuan et al., 2024), we formulate seven categories of personas: personality, beliefs, tastes, relationship, attributes, goals, and experience. See category details § A.1 and annotation examples in § A.2. Coarse-grained categories alone are insufficient to capture specific,

missing relevant persona descriptions. However, allowing fully free-form text makes it hard to determine whether two descriptions refer to the same underlying persona. To balance structure and expressiveness, we developed ten lexico-syntactic templates per category (*e.g.*, "*whether s/he (dis)likes VP*," where VP stands for a verb phrase), based on patterns observed in preliminary annotations. Refer to the full list of templates in Appendix A.3. We then describe the process of annotating missing relevant persona descriptions, which comprises three steps: (1) anonymization, (2) query-focused summarization, and (3) persona annotation.

Anonymization To ensure that persona identification relies solely on the provided basic persona description rather than human annotators' prior knowledge or models' parametric knowledge, we anonymize the dialogue data. Specifically, we replace all character names, organizations, geopolitical entities, facilities, and locations with placeholders (*e.g.*, "$person_1$," "org_1") using a named entity recognition (NER) following Sang et al. (2022).

Query-Focused Summarization Next, we produce a self-contained summary that clarifies the choice-making situation, ensuring that subsequent persona annotations are grounded in the same interpretation of the context. We ask three annotators to summarize each PICQ along with its preceding dialogue context in one sentence, as one example shown in Figure 1, and the instruction is shown in Appendix A.3. Each annotator works on two-thirds of the data, enabling overlap and cross-validation. The average agreement rate between annotator pairs is 88%, and consistency is judged by a natural language inference (NLI) model[1]. Disagreements are resolved through discussion.

Missing Relevant Persona Annotation Given the PICQ, context, summary, and the basic persona description, annotators are instructed to identify up to five missing persona descriptions most likely to influence the answer. For each, they first select a persona category from our predefined set (*e.g.*, Goal) and then describe the persona using a specific linguistic pattern associated with that category (Refer to Appendix A.3 for category details). Multiple personas per category are allowed, and irrelevant categories may be skipped. Annotations should prioritize personas serving as strong motivations, necessary conditions, and critical factors behind the choice when there are more than five relevant personas. Annotators are encouraged to generalize personas without losing core meaning and to avoid overly specific phrasing (*e.g.*, "*whether he likes Indian chicken curry*" to "*whether he likes curry*").

Each annotator labels two-thirds of the data to ensure overlap. Given that identifying relevant personas involves subjective judgment, prioritization, and potentially incomplete enumeration of personas, this overlap is designed to assess inter-annotator agreement and ensure annotation reliability. We automatically evaluate persona alignment using category matching and an NLI model.[1] Persona descriptions are considered to refer to the same persona if the NLI model predicts either entailment or contradiction (*e.g.*, "*whether he is introverted*" and "*whether he is extroverted*" are treated as aligned). On average, 59% of each annotator's persona annotations overlap with those of others.[2] The non-overlapping cases are partly attributed to the subjective nature of the task and the annotators' differing background knowledge and lived experiences, which may influence what aspects of the persona they perceive as relevant (*e.g.*, for a spending-related decision, annotators with different economic backgrounds may disagree on whether "*whether his financial status is good*" is relevant).

For the final dataset, we prioritize personas that are annotated by multiple annotators, keeping only one instance for each agreed-upon persona. If fewer than five such agreed-upon personas are available, we supplement them with additional non-overlapping ones. If more than five candidate personas exist, the annotators select the five most salient ones through discussion and reconciliation. Refer to Appendix A.3 for the annotation instructions. The dataset statistics are shown in Table 1.

Dialogue scenes w/ PICQ-answer pairs	289
PICQ-answer pairs influenced by persona	300
Total number of characters as listeners	60
Ave. number of utterances in dialogue context	22.69
Ave. number of tokens per utterance	17.23
Ave. identified relevant personas (after merging)	3.53
Ave. number of tokens per relevant persona	6.58

Table 1: Dataset PICQ-drama: statistics.

[1] https://huggingface.co/cross-encoder/nli-deberta-v3-base

[2] We recruit another external annotator to confirm the solidness of our annotation. Reading only the instructions, the annotator achieved 50% overlap with the annotators on 10% of the instances, demonstrating the task's reproducibility.

5 Evaluation of Identifying missing Relevant Persona

To empirically evaluate the task of identifying missing relevant personas, we conduct a series of experiments using LLMs (§ 5.1) to identify missing relevant persona and evaluating the indentified persona via various metrics (§ 5.2). Specifically, we answer the three research questions; RQ1: How well can leading LLMs identify missing relevant personas from context? (§ 5.3); RQ2: What cognitive patterns or differences emerge in their performance, especially compared to humans? (§ 5.4); RQ3: Can effective instruction strategies enhance model performance on this task? (§ 5.5).

5.1 Models

We evaluate closed- and open-source LLMs covering a range of architectures and scales. We used GPT-4.1-2025-04-14 as a strong closed-source LLM. We hereafter refer to it as GPT-4.1. We used Llama3.1-8B-Instruction,[3] Llama3.1-70B-Instruction,[4] Qwen3-8B,[5] and Qwen3-32B[6] models as open-source LLMs.

Instruction Strategies We test two types of instructions to solve the task, including one proposed for this task. **Baseline Prompting** directly asks the model to perform the identification task. In contrast, our proposed **Multi-task Prompting** first instructs the model to summarize the choice situation before identifying the personas, aiming to improve the comprehension, and then instructs the model to generalize the personas after the identification, aiming to avoid over-specific outputs. Refer to the prompts in Appendix A.4.

Human (Oracle) The annotations in our PICQ-drama dataset, listed and merged by the two human annotators for each query, serve as the ground truth for comparison across all metrics.

5.2 Metrics

We design a multi-faceted evaluation scheme to provide an assessment of model performance:

Influence measures the perceived impact of a missing persona on a character's decision. We use a 3-point Likert scale (0: irrelevant, 1: minor, 2: key). This metric reflects a model's ability to provide in-depth insights. Based on the manually summarized choice situation (§ 4.3), each persona is scored on whether it is irrelevant to the choice, slightly shapes preferences, or serves as a central motivation or constraint. We use GPT-4.1 (Temperature = 0.7) for full-scale automatic scoring. To validate this approach, we compared its ratings on 100 samples against human annotations, achieving a Cohen's κ of 0.658. The human inter-annotator agreement on the same set reached a κ of 0.831. Refer to Appendix A.5 for detailed definitions and the prompt.

Inaccessibility measures the difficulty of acquiring a missing persona, which also serves as a proxy for its privacy level. We use a 3-point Likert scale (0: Very Easy, 1: Easy, 2: Hard), where scores reflect whether a persona is observable by strangers (*e.g.*, gender), known to acquaintances, or requires close friendship. We use GPT-4.1 for full-scale automatic scoring. To validate this method, we compared its ratings on 100 samples with human annotations, achieving a Cohen's κ of 0.501. The inter-annotator agreement between two humans on the same set was a substantial κ of 0.618. Refer to Appendix A.5 for the detailed prompt.

Fidelity measures the semantic alignment between model-generated personas and the human-annotated gold references. The primary challenge is that free-text descriptions can be lexically different yet refer to the same underlying semantic factor. To handle this, we employ a two-step matching criterion: two personas are considered a match if (1) they belong to the same predefined persona category (*e.g.*, Beliefs), and (2) both descriptions probe the same persona with NLI (§ 4.3), persona descriptions are considered to refer to the same persona if the NLI model predicts either entailment or contradiction (*e.g.*, "whether he likes X" and "whether he dislikes X"). Based on this matching logic, we compute standard Precision, Recall, and F_1 scores to quantify the fidelity of a model's output.

Average number of missing relevant personas (Ave. Per.) is reported to show how many pieces of missing relevant personas the model identifies; we calculate the average number of missing relevant personas generated per PICQ of the models.

We perform paired bootstrap significance testing ($\alpha = 0.05$) on Influence, Inaccessibility, and F_1 scores to ensure that the claims in § 5.3 and § 5.5 are statistically significant.

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[5] https://huggingface.co/Qwen/Qwen3-8B

[6] https://huggingface.co/Qwen/Qwen3-32B

Models	Influence	Inaccessibility ↓	Fidelity			Ave. Per.
			Precision	Recall	F_1	
Llama3.1-8B	1.238	**1.385**	0.318	0.536	0.399	5.00
Llama3.1-8B-Multi	1.397	1.395	0.390	0.450	0.418	4.06
Llama3.1-70B	1.525	1.504	0.271	0.359	0.309	4.65
Llama3.1-70B-Multi	1.621	1.430	0.318	0.380	0.346	4.22
Qwen3-8B	1.377	1.485	0.345	0.456	0.393	4.66
Qwen3-8B-Multi	1.567	1.608	0.350	0.442	0.391	4.45
Qwen3-32B	1.510	1.558	0.399	0.567	0.468	5.00
Qwen3-32B-Multi	1.589	1.513	**0.409**	**0.581**	**0.480**	4.99
GPT-4.1	1.711	1.540	0.333	0.425	0.374	4.49
GPT-4.1-Multi	**1.772**	1.571	0.306	0.294	0.300	3.38
Human (Oracle)	1.648	1.394	–	–	–	3.53

Table 2: Results of identifying missing relevant personas.

5.3 Main Results

Table 2 shows the main results of our experiments. A key observation is that no single model excels across all metrics. Instead, the results reveal a complex, scale-dependent relationship between a model's ability to imitate human patterns (Fidelity) and its ability to generate profound analytical explanations (Influence).

Focusing on the fidelity dimension, as measured by the F_1 score, Qwen3-32B and Qwen3-32B-Multi achieve the highest Recall and F_1 scores, indicating that their generated personas align most closely with our human-annotated ground truth. In contrast, when evaluating for influence, as measured by the Influence score, GPT-4.1-Multi stands out, achieving the top score of 1.772. Most notably, this score surpasses even the Human (Oracle) baseline of 1.648. This suggests that GPT-4.1, when guided by our multi-task prompt, can identify personas perceived as even more impactful or fundamental to the decision than those articulated by humans.

Observing the scaling dynamics of fidelity and influence, we find that fidelity follows an inverted U-shaped trend. As model size increases from small (*e.g.*, Llama3.1-8B) to medium (*e.g.*, Qwen3-32B), fidelity improves. However, it declines for the largest models (e.g., Llama3.1-70B). In contrast, influence consistently increases with model size. Larger models are better at inferring missing personas, but as models become sufficiently large, their predictions become less aligned with human judgment patterns.

The Inaccessibility metric provides another critical layer of analysis. The Human (Oracle) exhibits the near lowest inaccessibility score (1.394), demonstrating remarkable efficiency. This aligns with the principle of "cognitive economy" in psychology (Rescher, 2017): humans are exceptionally skilled at identifying highly accessible (i.e., low inaccessibility) personas that still possess strong explanatory power (i.e., high influence). While some models like Llama3.1-8B achieve low inaccessibility, they do so at the cost of significantly lower influence.

Finally, the impact of our multi-task instruction strategy is consistent across the board. It systematically boosts the influence score for every model family, while also consistently reducing the Average Personas (Ave. Per.) generated. This confirms its role in encouraging models to perform analytical synthesis rather than simple enumeration. Some generated examples are shown in Appendix A.2.

5.4 Analysis of Cognitive Differences

Our results not only quantify model performance but also provide a window into the distinct cognitive models of different intelligent agents. We first diagnose the fundamental differences in attribution patterns between humans and LLMs. We then uncover the unique cognitive strengths of each agent type. Finally, based on these insights, we propose a synergistic framework that leverages these complementary advantages.

Table 3 shows the persona category distributions, revealing that humans and LLMs operate with fundamentally different cognitive patterns. Human annotators exhibit a clear cognitive model grounded in social context, heavily favoring Personality (28.9%), Taste (19.3%), and Relationship (25.2%). The most significant difference between humans and LLMs lies in two categories: humans prioritize Relationship, while LLMs, as a group,

Model	Personality	Belief	Taste	Relationship	Attribute	Goal	Experience
Llama3.1-8B	**28.9**	11.4	18.1	3.4	**20.6**	11.0	6.6
Llama3.1-70B	19.1	1.2	15.2	2.8	**40.3**	**21.0**	0.2
Qwen3-8B	**21.3**	17.7	**22.9**	5.8	14.0	18.1	0.3
Qwen3-32B	**26.9**	10.5	19.2	14.2	12.9	14.9	1.3
GPT-4.1	16.8	**21.4**	**27.9**	5.1	12.1	11.2	5.5
Human (Oracle)	**28.9**	14.3	19.3	**25.2**	6.0	4.2	2.3

Table 3: Distribution of Identified Relevant Persona Types (%). (percentages higher than 20% are bold).

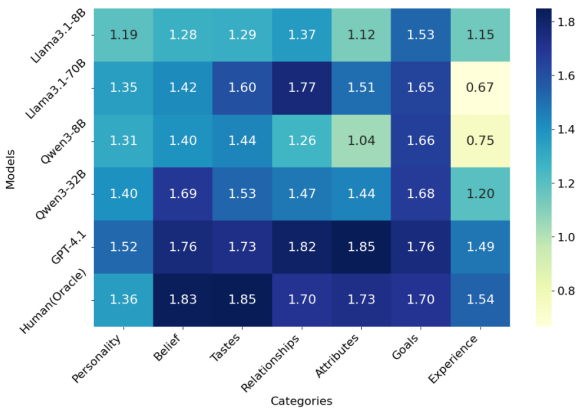

Figure 3: Influence heatmap of identified personas according to categories.

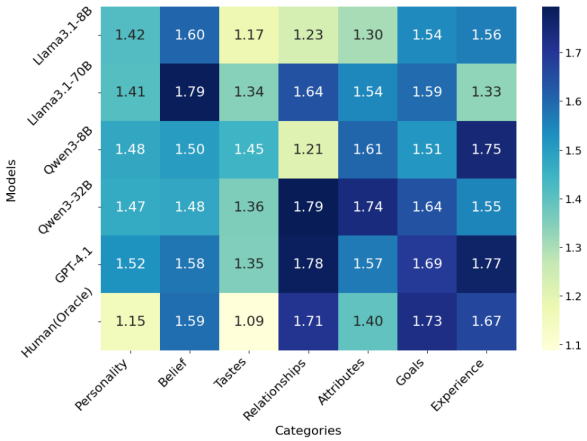

Figure 4: Inaccessibility heatmap of identified personas according to categories.

consistently favor Goal. This points to a core divergence: human reasoning is deeply embedded in social dynamics, whereas LLMs appear to operate under a more utilitarian, task-oriented framework, assuming a goal-driven motive behind actions. While LLMs share this general tendency, there are important outliers. Qwen3-32B, with its relatively high emphasis on Relationship, stands out as the most "human-like" LLM, explaining its top-tier fidelity. In contrast, the Llama series displays a unique bias towards Attribute, offering a different perspective on choice-making.

Figures 3 and 4 allow us to move to identifying the characteristics of each persona category and the unique strengths of each agent. The categories themselves exhibit distinct profiles. Personality acts as a global trait with moderate influence and low accessibility. Belief, Goal, and Relationship function as deep motivators, being both high-impact and hard to acquire. Taste serves as a direct driver, being highly influential and easily accessible. We can observe that the GPT-4.1 model achieves consistently high Influence across nearly all categories, capable of uncovering high-impact motives. While humans excel at identifying personas in the Personality, Taste, and Attribute categories that yield extremely high influence for a very low inaccessibility cost. This aligns with the principle of "cognitive economy," positioning humans as experts in finding high-yield, low-effort explanations.

Our analysis suggests that the path forward is not selecting a single best agent, but combining complementary persona sources into a synergistic framework for identifying missing relevant personas. For dataset construction, our task can guide developers to collect minimal yet sufficient personas, improving simulation fidelity. This can be implemented as a three-stage process: divergent persona generation, convergent filtering to remove redundancy, and final human selection to ensure relevance. For simulating a specific individual, the framework can operate as a persona completion process, where a model identifies missing persona dimensions for the current scenario and prompts the user (or another module) to provide them.

5.5 Ablation Study

To understand the contribution of each component in our multi-task instruction strategy, we conduct an ablation study. Our strategy combines two steps: summarizing the choice situation and generalizing the identified personas. We analyze the impact

Models	Influence	Inaccessibility ↓	Fidelity			Ave. Per.
			Precision	Recall	F_1	
Qwen3-32B-Multi	1.589	**1.513**	0.409	0.581	0.480	4.99
− Summarization	1.500	1.541	0.411	0.564	0.476	4.83
− Generalization	**1.594**	1.556	0.410	0.590	**0.484**	5.00
− Summarization & Generalization	1.510	1.558	0.399	0.567	0.468	5.00
GPT-4.1-Multi	1.772	1.571	0.306	0.294	0.300	3.38
− Summarization	**1.794**	**1.538**	0.347	0.274	0.306	2.78
− Generalization	1.698	1.556	0.316	0.430	0.364	4.79
− Summarization & Generalization	1.711	1.540	0.333	0.425	**0.374**	4.49

Table 4: Ablation results of our multi-task instruction strategy.

of these components on two models that have the best performance on influence and fidelity aspects: GPT-4.1 and Qwen3-32B.

Table 4 shows the results. The Summarize component yields a significant improvement only for Qwen3-32B in terms of Influence. This supports its intended role of helping the model correctly interpret the specific choice situation. For GPT-4.1, no significant gain is observed, suggesting its baseline comprehension is already sufficient. In contrast, the Generalize component has a significant effect only on GPT-4.1, increasing influence while reducing fidelity. However, the resulting abstraction exceeds the level of cognitive effort typically exercised by human annotators, leading to aggressive merging of persona dimensions, fewer generated personas, and lower fidelity to human patterns. Overall, the Summarize component benefits mid-sized models by improving the understanding of the choice situation, while the Generalize component primarily affects large models by triggering deeper abstraction, increasing influence at the cost of fidelity.

6 Conclusions

This work highlights a common oversight in user simulation: the assumption that provided persona information is sufficient. We formalize this as a new task, identifying missing relevant persona dimensions, and present the first benchmark for its evaluation. Using our PICQ-drama dataset, we demonstrate the feasibility of applying LLMs to this task. Our results show that the ability to detect influential missing personas generally increases with model scale. The discovery of an inverted U-shaped fidelity curve, linked to the concept of human "cognitive economy," offers a novel lens for comparing human and LLM cognition. Our further analysis shows the cognitive differences within LLMs, as well as between LLMs and humans. Besides, we design a multi-task instruction strategy that improves the LLMs' ability to identify missing personas that better influence the choices.

Future work will focus on collecting the identified missing personas to evaluate their direct impact on the downstream simulation tasks.

Acknowledgement

This work was supported by Institute for Digital Observatory, the University of Tokyo.

Limitations

Our study is limited to English-language data. Differences in language and sociocultural background, whether in model training or human annotation, may lead to divergent interpretations of what constitutes a relevant persona. As such, the identified personas and their perceived influence on user responses may vary across linguistic and cultural contexts, suggesting that further exploration is needed to understand and generalize these findings across languages and cultures.

Our study is based on drama scripts rather than spontaneous, real-world conversations. This choice was made primarily to navigate the significant ethical challenges, such as privacy and consent, associated with collecting and analyzing authentic personal dialogues. To maximize the resemblance to reality, we selected scripts from sitcoms and dramas that focus on everyday life, interpersonal relationships, and common choice-making scenarios. However, an unavoidable gap remains. Scripted dialogue is typically more structured, coherent, and sometimes theatrically heightened compared to authentic speech, which is often disfluent and fragmented. Future research should aim to validate and extend our findings on datasets of anonymized, ethically-sourced real-world conversations to assess the generalizability of our findings.

References

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Trans. Mach. Learn. Res.*, 2024.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond demographics: Aligning role-playing LLM-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, Miami, Florida, USA. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *Preprint*, arXiv:2307.14984.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

LlamaTeam. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Man Tik Ng, Hui Tung Tse, Jen tse Huang, Jingjing Li, Wenxuan Wang, and Michael R. Lyu. 2024. How well can llms echo us? evaluating ai chatbots' role-play ability with echo. *Preprint*, arXiv:2404.13957.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Nicholas Rescher. 2017. *Cognitive economy: The economic dimension of the theory of knowledge*. University of Pittsburgh Pre.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character comprehension in stories as speaker guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *Preprint*, arXiv:2501.15427.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Categories of Personas

- **Personality** Stable psychological traits such as extroversion, or emotional sensitivity. These traits influence how a person tends to behave or react in various situations.
- **Beliefs** Enduring convictions or values, including moral principles, or political stances, shaping a person's judgment of what is right.
- **Tastes** Personal preferences for things like food, music, or entertainment. Tastes can strongly affect choices involving consumption, participation, or lifestyle.
- **Relationship** Social ties and interpersonal history with other characters, such as being a friend, sibling, or coworker. These relationships influence the level of trust, obligation, or emotional support, which can significantly shape one's choice.
- **Attributes** Basic biographical or demographic characteristics, such as income level, occupation, or cultural background. These factors may constrain or inform choices due to physical capability or role expectations
- **Goals** Current intentions, needs, or objectives a person is trying to achieve. Goals directly impact choices by framing what is desirable or prioritized in a given situation.
- **Experience** Experience would be chosen only if the specific past event memory itself is directly influencing the choice, rather than having been internalized in other categories.

A.2 Generated Examples

Tables 5 and 6 some generated examples, we picked GPT-4.1-Multi (high influence) and Qwen3-32-Multi (high fidelity) settings, and human annotations. These examples highlight the different tendencies of models. GPT-4.1-Multi tends to generate more concise and impactful explanations, whereas Qwen3-32B-Multi produces more constrained, context-grounded personas.

A.3 Instructions for Annotation

Table 8 shows the instructions for human annotators to check potential PICQs and their corresponding answers. Table 9 shows the instructions for human annotators to write query-focused summarization. Table 10 shows the instructions for human annotators to annotate the missing relevant personas

A.4 Prompts for Identifying PICQs and missing Relevant Personas

Table 7 shows the prompts for LLMs to detect potential PICQs and their corresponding answers. Table 11 shows the prompts for LLMs to identify the missing relevant personas. Table 12 shows the prompts for LLMs to identify the missing relevant personas using a multi-task instruction strategy.

A.5 Prompts for LLM-as-Judge

Table 13 shows the prompts for LLMs to evaluate the Influence score of the identified missing relevant personas. Table 14 shows the prompts for LLMs to evaluate the inaccessibility score of the identified missing relevant personas.

[Example 1] Context & Identified Missing Relevant Persona Dimensions

Context: Background Scene 6: Central Perk (Rachel Is Serving Coffee. Enter Mindy)
(2) Rachel: Mindy.
(3) Mindy: Hey, you.
(4) Rachel: Hey, you.... So, what's up?
(5) Mindy: Um.. we should really be sitting for this.
(6) Rachel: Sure we should... So.
(7) Mindy: Now- I know things've been weird lately, but you're like my oldest friend in the world.. except for maybe Laurie Schaffer, who I don't talk to anywhere, 'cause she's all bitter now that she lost the weight and it turns out she doesn't have a pretty face. ....Okay, I'm just gonna ask you this once, and I want a straight answer.
(8) Rachel: Okay.
(9) Mindy: **Will you be my maid of honor?**

GPT-4.1-Multi
- **tastes**: whether she likes or dislikes being involved in major ceremonial roles such as maid of honour
- **personality**: whether she is outgoing and enjoys social responsibility
- **relationships**: whether she feels emotionally connected to Mindy at the present time
- **experience**: whether she has had past experiences that shape her willingness to accept this role
- **attributes**: whether she has enough free time or flexibility in her personal schedule to take this on

Qwen3-32B-Multi
- **beliefs**: whether she believes it is important to support close friends in significant life events
- **personality**: whether she is reliable
- **relationships**: whether she is loyal to Mindy
- **goals**: whether she aims to maintain a strong friendship with Mindy
- **personality**: whether she is comfortable with taking on responsibilities

Human
- **relationships**: whether she is close to Mindy
- **tastes**: whether she likes attending wedding ceremonies
- **personality**: whether she is helpful
- **personality**: whether she is shy in public
- **goals**: whether she aims to be a maid of honor

Table 5: Generated example 1.

[Example 2] Context & Identified Missing Relevant Persona Dimensions

Context: Background ([SCENE_BREAK])
(1) Michael: I should have never let the Temp touch this thing. I had all these great icons and now I have four folders. So..
(2) Dwight: It's actually better this way.
(3) Michael: No it's not. Because I could just click on the icon and then I'm onto—
(4) Dwight: Michael, could I ask you something? I wanted to ask your permission to ask out Katy. I know it's against the rules and everything. Because...
(5) Michael: No, no, no it's not against the rules. She's not a permanent employee so it's not.
(6) Dwight: Thank you, Michael. I appreciate this so much.
(7) Michael: But I think you should just know that I am going to be giving her a ride home later.
(8) Dwight: What?
(9) Michael: She asked me for a ride and so I am going to give her a ride home.
(10) Dwight: Is that all it is? Just a ride home? Like a taxicab?
(11) Michael: Well, might be a ride home. Might be a ride home and we stop for coffee and dot-dot-dot...
(12) Dwight: Please. Please, I am your inferior and I'm asking you this favor. **Can you promise me that it will just be a ride home?**

GPT-4.1-Multi
- **beliefs**: whether he values professional ethics in interpersonal relationships
- **personality**: whether he is empathetic or enjoys holding power in ambiguous situations

Qwen3-32B-Multi
- **personality**: whether he is trustworthy
- **personality**: whether he is considerate
- **beliefs**: whether he believes it's important to respect others' feelings
- **relationships**: whether he is friendly with Katy
- **goals**: whether he aims to maintain a professional relationship with Dwight

Human
- **relationship**: whether he is fond of Katy
- **relationship**: whether he is close with Dwight
- **beliefs**: whether he believes it's fine to get an employee to be his girlfriend
- **attributes**: whether his pattern status is 'in a relationship'
- **personality**: whether he is gentle

Table 6: Generated example 2.

```
Please review the given dialogue and identify all specific question-answering pairs that consist
of:
  - An utterance that asks a specific listener to choose among specific options whose choice will
implicitly depend on their personas (e.g., personality traits, beliefs, preference, relationships,
attributes, goals, and experiences).
  - A response by the listener which makes a specific choice among the given options in the format
specified after ### Format.

Please solve the above task step by step as follows:
Step 1. Find the next question utterance in the given dialogue, and continue to Step 2 if found;
otherwise, exit.
Step 2. Judge whether the question found in Step 1 asks a single listener to choose among specific
options, continue to Step 3 if yes; otherwise, go to Step 1.
  - Questions without specific options being presented should be excluded (e.g., "where are you
going?")
Step 3. Judge whether the choice would be influenced by the listener's personas, continue to Step 4
if yes; otherwise, go to Step 1.
  - The question that directly asks a piece of persona should be excluded (e.g., "Do you like
apples?")
  - The question that directly asks a fact should be excluded (e.g., "Did you get the cookie?")
Step 4. Find the nearest response by the listener in which the listener makes a specific choice
among the given options, return the two utterances as pairs if found; otherwise, go to Step 1.

### Format:
Each utterance in the dialogue has a unique numeric identifier (e.g., 1, 2, 3...).
Return your result as a list of **(question, response)** number pairs:
[(3, 5), (8, 9), ...]
```

Table 7: Prompt for detecting potential PICQs and their corresponding answers.

Please review the specific question-answering pairs identified by GPT-4.1 and the whole scene dialogue, then judge whether each pair is correct or incorrect based on the guidelines below.

Task:
For each pair (question utterance, answer) determine whether:
1. Whether the question asks a single listener to choose among specific options
   a. Question without specific options is incorrect (e.g., "where are you going?")
2. Whether the choice would be influenced by the listener's personas (e.g., personality traits, beliefs, preference, relationships, attributes, goals, and experiences)
  a. Question that directly asks a piece of persona incorrect (e.g., "Do you like apples?")
  b. Question that directly asks a fact is incorrect (e.g., "Did you get the cookie?")
3. Whether the answer is the nearest response by the listener in which the listener makes a specific choice among the given options

If the pair satisfies all the above requirements, label correct, otherwise incorrect.

Table 8: Instruction for human annotators to check potential PICQs and their corresponding answers.

Given dialogue context, question utterance, and basic persona (age, gender, and basic relationship with the questioner) of the listener, your task is to:

Conduct query-focused summarization based on the question and dialogue context. The summary should be self-contained, allowing someone to understand the listener's choice-making situation (what they need to choose and why) just by reading the summary, without referring back to the original dialogue.
  a. If the dialogue context and question do not contain enough specific information about the listener's choice-making situation (or are not self-contained), skip them.

Table 9: Instruction for query-focused summarization.

Given dialogue context, question utterance, and basic persona (age, gender, and basic relationship with the questioner) of the listener, your task is to:

Identify missing persona (not explicitly stated anywhere in the provided context and basic persona) of the listener that will influence the choice of the options provided in the question the most.
a. You should identify up to five pieces of missing persona. Prioritize based on factors that:
i. Represent strong motivations or driving forces behind the choice (e.g., key personal goals, deeply held beliefs, very strong preferences).
ii. Act as necessary conditions, core constraints, or essential enablers (e.g., affordability for a large purchase, prerequisite required skills).
iii. Are critical factors when their value falls within a certain range (e.g., “spice tolerance” when choosing a Sichuan (spicy) vs. Japanese restaurant).
b. When you consider each piece of the required persona, you should first choose a category and choose the specific linguistic patterns associated with the category to describe the specific persona required to make a choice. You can write more than one piece of the persona for the same category, and you can skip some categories if they are irrelevant. We count the number of pieces of the required persona by the number of descriptions you have provided (different pieces of persona in the same category should be treated):
i. Personality:
1. whether s/he is ADJ (ADJ is an adjective describing a personality trait); for example,
- whether s/he is introverted
- whether s/he is adventurous
ii. Beliefs (personal values, moral principles, and views on social norms):
1. whether s/he believes it’s ADJ to VP (ADJ is an adjective to comment a behavior, VP is a verb phrase); for example,
- whether s/he believes it’s important to save money
- whether s/he believes it’s wrong to lie to others
2. whether s/he believes propN should VP (propN is a target, VP is a verb phrase); for example,
- whether s/he believes children should have less screen time
- whether s/he believes the government should invest more in public transport
iii. Tastes:
1. whether s/he (dis)likes VP (VP is a verb phrase); for example,
- whether s/he likes traveling
- whether s/he dislikes waking up early
2. whether s/he (dis)likes N (N is a noun); for example,
- whether s/he likes spicy food
- whether s/he dislikes crowded places
iv. Relationships:
1. whether s/he is N of propN (N is a noun representing a human relationship, propN is a name of a person); for example,
- whether s/he is a close friend of Alex
- whether s/he is the sibling of Sarah
2. whether s/he is ADJ + P + propN (ADJ represents an adjective or a past participle used adjectivally, describes the subject’s (s/he’s) view, attitude, feeling, or judgment regarding the person propN, P is prepositional); for example,
- whether s/he is annoyed with Maria
- whether s/he is loyal to their team
v. Attributes:
1. whether his/her ATTR is X (ATTR is a noun describing an attribute of a person, (e.g., gender, occupation, age, height, weight, income, etc.) X is a specific attribute value or an expression describing a range of values); for example,
- whether his/her physical stamina is suitable for a long hike
- whether his/her disposable income is suitable for luxury purchases
vi. Goals (short-term or long-term goals):
1. whether s/he aims to VP (VP is a verb phrase describing the goal); for example,
- whether s/he aims to get a promotion
- whether s/he aims to learn a new language
vii. Experience (Write ’experience’ only if the specific past event memory itself is directly influencing the choice, rather than having been internalized as a taste or other categories):
1. whether s/he has V (V is a past participle phrase); for example,
- whether s/he has been to that restaurant before
- whether s/he has had a bad experience with online shopping
c. After identifying a specific piece of persona that significantly influences the choice, consider if it can be expressed in a more generalized or abstract way without losing its core impact or clarity. Meanwhile, prioritize annotating plausible and impactful missing personas, avoiding overly specific or highly improbable scenarios unless contextually supported; for example,
- whether s/he likes Indian chicken curry -> whether s/he likes curry
- whether s/he is helpful in park at night -> whether s/he is helpful
d. Apart from using "s/he" to refer to the respondent, do not use pronominal reference for other entities, even if they are present in the context.

Table 10: Instruction for identifying missing relevant personas.

```
Given dialogue context, question utterance, and basic persona (age, gender, and basic relationship with the questioner) of the
listener, your task is to:

Identify missing persona (not explicitly stated anywhere in the provided context and basic persona) of the listener that will
influence the choice of the options provided in the question the most.
   a. You should identify up to five pieces of missing persona. Prioritize based on factors that:
    i. Represent strong motivations or driving forces behind the choice (e.g., key personal goals, deeply held beliefs, very
strong preferences).
    ii.  Act as necessary conditions, core constraints, or essential enablers (e.g., affordability for a large purchase,
prerequisite required skills).
    iii. Are critical factors when their value falls within a certain range (e.g., “spice tolerance” when choosing a Sichuan
(spicy) vs. Japanese restaurant).
   b. When you consider each piece of the required persona, you should first choose a category and choose the specific linguistic
patterns associated with the category to describe the specific persona required to make a choice. You can write more than one
piece of the persona for the same category, and you can skip some categories if they are irrelevant. We count the number of
pieces of the required persona by the number of descriptions you have provided (different pieces of persona in the same category
should be treated):
    i. Personality:
      1. whether s/he is ADJ (ADJ is an adjective describing a personality trait); for example,
        - whether s/he is introverted
        - whether s/he is adventurous
    ii. Beliefs (personal values, moral principles, and views on social norms):
      1. whether s/he believes it’s ADJ to VP (ADJ is an adjective to comment a behavior, VP is a verb phrase); for example,
        - whether s/he believes it’s important to save money
        - whether s/he believes it’s wrong to lie to others
      2. whether s/he believes propN should VP (propN is a target, VP is a verb phrase); for example,
        - whether s/he believes children should have less screen time
        - whether s/he believes the government should invest more in public transport
    iii. Tastes:
      1. whether s/he (dis)likes VP (VP is a verb phrase); for example,
        - whether s/he likes traveling
        - whether s/he dislikes waking up early
      2. whether s/he (dis)likes N (N is a noun); for example,
        - whether s/he likes spicy food
        - whether s/he dislikes crowded places
    iv. Relationships:
      1. whether s/he is N of propN (N is a noun representing a human relationship, propN is a name of a person); for example,
        - whether s/he is a close friend of Alex
        - whether s/he is the sibling of Sarah
      2.  whether s/he is ADJ + P + propN (ADJ represents an adjective or a past participle used adjectivally, describes the
subject’s (s/he’s) view, attitude, feeling, or judgment regarding the person propN, P is prepositional); for example,
        - whether s/he is annoyed with Maria
        - whether s/he is loyal to their team
    v. Attributes:
      1. whether his/her ATTR is X (ATTR is a noun describing an attribute of a person, (e.g., gender, occupation, age, height,
weight, income, etc.) X is a specific attribute value or an expression describing a range of values); for example,
        - whether his/her physical stamina is suitable for a long hike
        - whether his/her disposable income is suitable for luxury purchases
    vi. Goals (short-term or long-term goals):
      1. whether s/he aims to VP (VP is a verb phrase describing the goal); for example,
        - whether s/he aims to get a promotion
        - whether s/he aims to learn a new language
    vii. Experience (Write ’experience’ only if the specific past event memory itself is directly influencing the choice, rather
than having been internalized as a taste or other categories):
      1. whether s/he has V (V is a past participle phrase); for example,
        - whether s/he has been to that restaurant before
        - whether s/he has had a bad experience with online shopping
Output strictly in the following format:
(personality) whether he is introverted
(tastes) whether she dislikes waking up early.

Do not output additional explanation!
Here is the basic persona about {choice-maker}: {basic_info}
The following is the conversation, with the final utterance being the question utterance:
```

Table 11: Prompt for identifying missing relevant personas.

```
Given dialogue context, question utterance, and basic persona (age, gender, and basic relationship with the questioner) of the
listener, your task is to:

Identify missing persona (not explicitly stated anywhere in the provided context and basic persona) of the listener that will
influence the choice of the options provided in the question most.
   a. You should identify up to five pieces of missing persona. Prioritize based on factors that:
    i. Represent strong motivations or driving forces behind the choice (e.g., key personal goals, deeply held beliefs, very
strong preferences).
    ii.  Act as necessary conditions, core constraints, or essential enablers (e.g., affordability for a large purchase,
prerequisite required skills).
    iii. Are critical factors when their value falls within a certain range (e.g., “spice tolerance” when choosing a Sichuan
(spicy) vs. Japanese restaurant).
   b. When you consider each piece of the required persona, you should first choose a category and choose the specific linguistic
patterns associated with the category to describe the specific persona required to make a choice. You can write more than one
piece of the persona for the same category, and you can skip some categories if they are irrelevant. We count the number of
pieces of the required persona by the number of descriptions you have provided (different pieces of persona in the same category
should be treated):
    i. Personality:
      1. whether s/he is ADJ (ADJ is an adjective describing a personality trait); for example,
        - whether s/he is introverted
        - whether s/he is adventurous
    ii. Beliefs (personal values, moral principles, and views on social norms):
      1. whether s/he believes it’s ADJ to VP (ADJ is an adjective to comment a behavior, VP is a verb phrase); for example,
        - whether s/he believes it’s important to save money
        - whether s/he believes it’s wrong to lie to others
      2. whether s/he believes propN should VP (propN is a target, VP is a verb phrase); for example,
        - whether s/he believes children should have less screen time
        - whether s/he believes the government should invest more in public transport
    iii. Tastes:
      1. whether s/he (dis)likes VP (VP is a verb phrase); for example,
        - whether s/he likes traveling
        - whether s/he dislikes waking up early
      2. whether s/he (dis)likes N (N is a noun); for example,
        - whether s/he likes spicy food
        - whether s/he dislikes crowded places
    iv. Relationships:
      1. whether s/he is N of propN (N is a noun representing a human relationship, propN is a name of a person); for example,
        - whether s/he is a close friend of Alex
        - whether s/he is the sibling of Sarah
      2. whether s/he is ADJ + P + propN (ADJ represents an adjective or a past participle used adjectivally, describes the
subject’s (s/he’s) view, attitude, feeling, or judgment regarding the person propN, P is prepositional); for example,
        - whether s/he is annoyed with Maria
        - whether s/he is loyal to their team
    v. Attributes:
      1. whether his/her ATTR is X (ATTR is a noun describing an attribute of a person, (e.g., gender, occupation, age, height,
weight, income, etc.) X is a specific attribute value or an expression describing a range of values); for example,
        - whether his/her physical stamina is suitable for a long hike
        - whether his/her disposable income is suitable for luxury purchases
    vi. Goals (short-term or long-term goals):
      1. whether s/he aims to VP (VP is a verb phrase describing the goal); for example,
        - whether s/he aims to get a promotion
        - whether s/he aims to learn a new language
    vii. Experience (Write ’experience’ only if the specific past event memory itself is directly influencing the choice, rather
than having been internalized as a taste or other categories):
      1. whether s/he has V (V is a past participle phrase); for example,
        - whether s/he has been to that restaurant before
        - whether s/he has had a bad experience with online shopping
    3. After identifying a specific piece of persona that significantly influences the decision, consider if it can be expressed
in a more generalized or abstract way without losing its core impact or clarity to avoid overly specific or highly improbable
scenarios unless contextually supported. For example,
      - whether s/he likes Indian chicken curry -> whether s/he likes curry
      - whether s/he is helpful in park at night -> whether s/he is helpful
    4. You should first summarize what choice maker is requested to make and then identify the missing persona, finally generalize
them
Output strictly in the following format:
(summary) ...
(personality) whether he is introverted
(tastes) whether he likes Indian chicken curry
...
[generalized]
(personality) whether he is introverted
(tastes) whether he likes curry
...
Do not output additional explanation!
Here is the basic persona about {choice-maker}: {basic_info}
The following is the conversation, with the final utterance being the question utterance:
```

Table 12: Prompt for identifying missing personas in multi-task instruction strategy.

Given a summary showing the target's choice-making situation (what they need to choose and why) and their basic persona and a list of persona dimensions (meaning the specific value of that persona for the target is currently unknown), your task is to assign an influence score for each piece of persona dimension considering their possible influence on the choice-making.

Definition of influence score:

Influence refers to how strongly a piece of persona dimension affects the answer to the question. It is rated as follows:

Score 0 - Irrelevant
The persona dimension, regardless of its value or state, has no influence on the choice outcome.
e.g., "favorite color" generally has no impact on deciding whether to accept a job offer.

Score 1 - Minor Influence
The persona dimension has some influence on the choice-making process or outcome, but this influence is not strong enough to be considered key or decisive. It might affect preferences for details, execution, or make one option slightly more or less appealing, but it does not fundamentally drive or constrain the core choice.
e.g., "A slight preference for Software X's user interface over Software Y's, when both tools meet all core functional requirements and are within budget," might influence which tool the team adopts, but it would not lead them to choose Software X if it lacked a critical feature that Software Y possessed.

Score 2 - Key Influence
The persona dimension is a key factor that influences, changes, or determines the main choice outcome. This influence can manifest as a necessary condition/constraint/enabler (making an option impossible or essential under certain conditions) OR as a strong motivation that shapes the choice.
e.g., "Spice tolerance" is key if extremely low (effectively vetoes Sichuan, acting as a constraint). "Having a driver's license" is key if the job requires driving (a necessary condition).
e.g., "whether s/he likes alcohol" significantly influences a choice about drinking beer (as a strong motivation) whether s/he likes or dislikes alcohol.

You should first consider the possible values of each piece of missing personal information and then judge its score.

Summary, basic persona, and missing persona dimensions are given as follows:

Table 13: Prompt for calculating the influence score for missing relevant personas.

Your task is to analyze the persona dimensions (meaning the specific value of that persona for the target is currently unknown) from the perspective of the target individual based on the provided Basic Persona.
You will receive the persona dimensions belonging to that same individual.
You should assign an inaccessibility score (0-2) to each Info. This score reflects the individual's willingness to disclose that specific persona dimension, based on the benchmark of who is proactively asking them about it.

Scoring Scale:
Score 0: Public / Observable Information
Willingness Benchmark: The individual would share this with a stranger, or the information is physically observable/public knowledge anyway.
Examples: name, visible appearance (hair color, height), accent.

Score 1: General Acquaintance Information
Willingness Benchmark: The individual would NOT share this with a random stranger, but would comfortably share it with a general acquaintance (like a co-worker, a neighbor, or a casual friend) if the topic came up in conversation.
Examples: Job title, general hobbies, hometown, favorite sports team, where s/he went to college.

Score 2: Close Acquaintance Information
Willingness Benchmark: The individual would only share this information with a close acquaintance whom they trust (such as a close friend, immediate family, or a spouse). They would actively avoid discussing this with co-workers or casual friends.
Examples: specific salary/financial struggles, private family issues, detailed medical conditions, strong political opinions, deep life aspirations.

You may first imagine you are the individual and then consider whether it is comfortable for you to share the personal information during the conversation with a certain group of people.
Basic persona description and persona dimensions are given as follows:

Table 14: Prompt for calculating the inaccessibility score for missing relevant personas.