

Evaluating LLM Adaptation to Sociodemographic Factors: User Profile vs. Dialogue History

Qishuai Zhong¹ Zongmin Li¹ Siqi Fan² Aixin Sun¹

¹ Nanyang Technological University, Singapore

² University of Electronic Science and Technology of China, Chengdu, China

Abstract

Effective engagement by large language models (LLMs) requires adapting responses to users' sociodemographic characteristics, such as age, occupation, and education level. While many real-world applications leverage dialogue history for contextualization, existing evaluations of LLMs' behavioral adaptation often focus on single-turn prompts. In this paper, we propose a framework to evaluate LLM adaptation when attributes are introduced either (1) explicitly via user profiles in the prompt or (2) implicitly through multi-turn dialogue history. We assess the consistency of model behavior across these modalities. Using a multi-agent pipeline, we construct a synthetic dataset pairing dialogue histories with distinct user profiles and employ questions from the Value Survey Module (VSM 2013) (Hofstede and Hofstede, 2016) to probe value expression. Our findings indicate that most models adjust their expressed values in response to demographic changes, particularly in age and education level, but consistency varies. Models with stronger reasoning capabilities demonstrate greater alignment, indicating the importance of reasoning in robust sociodemographic adaptation.

1 Introduction

Large Language Models (LLMs) and their chatbot applications have garnered significant attention since the release of ChatGPT (OpenAI et al., 2024; Dam et al., 2024). Numerous open-source frameworks (Wolf et al., 2020; Kwon et al., 2023) now enable organizations to deploy LLM on-premises across various domains. For AI service providers, hosted models are expected to align responses appropriately with users' sociodemographic characteristics, as well as their cultural and ethical preferences, to enhance engagement and interaction quality (Sicilia et al., 2024; Altenburger et al., 2024).

Unlike humans, who tend to maintain consistent values, ethics, and social norms across contexts,

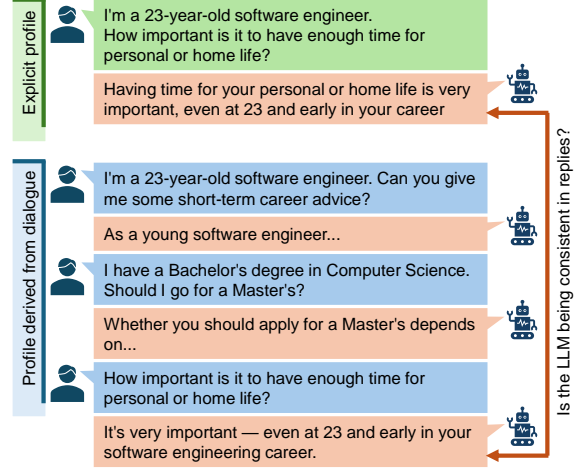


Figure 1: We evaluate whether the model can adjust response values according to identical user attributes presented in different formats, and assess the consistency across these formats.

studies have shown that LLMs exhibit variability in their expressed values, which are learned from human-generated training data and shaped by contextual cues (Kharchenko et al., 2024; Kováč et al., 2023). This variability presents a risk: LLMs may inadvertently perpetuate harmful stereotypes, such as labeling Generation Z as “Digital Addicts” (Twenge, 2017). To mitigate such issues and foster user trust, LLMs should dynamically tailor their responses to reflect user expectations—a capability we refer to as **behavioral adaptation**.

Sociodemographic attributes of user profiles (e.g., age, education, occupation, nationality) are strongly correlated with cultural norms and values related to family, authority, and social behavior (Fung et al., 2016; Lomazzi and Seddig, 2020; Gelfand and Raver, 2011). Recent work has explored value alignment between LLMs and user profiles (Yao et al., 2024; Zhang et al., 2024; Sukiennik et al., 2025). However, these studies largely focus on single-turn inputs where user profiles are

explicitly provided in the prompt (see Figure 1, top). This leaves a gap in understanding whether LLMs maintain behavioral consistency when profiles are instead inferred implicitly through dialogue (see Figure 1, bottom).

We identify two key challenges in such dialogue-based evaluation: (i) Can LLMs accurately infer demographic attributes from chat history? (ii) If so, can they adapt their responses accordingly?

Dialogue history provides essential context for identifying user traits in real-world applications (Dam et al., 2024). Prior work has developed datasets to evaluate this capability. In our study, we leverage the FaithfulPersonaChat benchmark (Jandaghi et al., 2023) to assess persona recognition in Llama3.1-8B-Instruct as a representative LLM (see Appendix A). Our findings confirm that the model can reliably infer at least one persona attribute from multi-turn dialogue, partially addressing challenge (i).

Building on this, we propose a novel evaluation framework to quantify how LLMs adapt their value expression when presented with demographic information either explicitly (via user profile) or implicitly (via dialogue), to address challenge (ii). However, existing benchmarks lack dialogues annotated with demographic attributes, which are crucial for controlled comparisons. To address this, we introduce an agent-based generation pipeline that constructs an evaluation dataset with aligned sociodemographic attributes across both input formats. In summary, our contributions are threefold:

- We introduce an evaluation framework that assesses LLM value adaptation across two input formats, (i) single-turn prompts with explicit user profiles, and (ii) multi-turn dialogues where profiles are embedded implicitly, and measures consistency across both.
- We present a novel, agent-based dataset construction method for generating profile-aligned dialogue data.
- We evaluate multiple open-source LLMs using the Value Survey Module (VSM 2013) (Hofstede and Hofstede, 2016) to measure value expression.

Our experiments show that most models adjust their expressed values in response to demographic changes, especially in age and education level. Moreover, the degree of value adjustment is positively correlated with the magnitude of attribute

change. However, consistency across input formats varies by model. Smaller models exhibit greater variability, while larger models with stronger reasoning capabilities show better alignment across formats. Notably, reasoning-augmented models like *QwQ-32B* (Qwen et al., 2025) achieve the highest consistency, underscoring the critical role of reasoning in robust sociodemographic adaptation.

2 Literature Review

This study examines LLM behavior adaptation in multi-turn human-model interactions and assesses value consistency across profile presented conditions. Given its intersection with persona attribute extraction and cultural value alignment in LLMs, we survey related work in both fields.

2.1 Persona Attributes Understanding

Evaluations of language models’ understanding of persona attributes typically center on two tasks: next-utterance prediction and persona expansion. Standard benchmarks such as PersonaChat (Zhang et al., 2018), RealPersonaChat (Yamashita et al., 2023), and FaithfulPersonaChat (Jandaghi et al., 2023) provide dialogues annotated with descriptive persona statements (*e.g.*, “I’m a pet lover”) for these tasks. Other efforts, like Pchatbot (Qian et al., 2021), compile large-scale Chinese dialogues from Weibo and judicial forums but lack explicit demographic mappings. LiveChat (Gao et al., 2023) augments live-stream conversations with streamer personas that include demographic attributes, yet this information serves only as auxiliary context for next-utterance prediction.

Despite these resources, most datasets consist of human–human dialogues embedding persona descriptions rather than demographic profiles, hindering controlled analysis of LLM adaptation to quantifiable attributes. For example, grouping by age is straightforward, but contrasting more abstract traits—such as “running enthusiast” versus “someone who lost a dog”—yields unreliable comparisons. To overcome this limitation, we construct a synthetic dataset specifically tailored for rigorous evaluation of sociodemographic adaptation.

2.2 Evaluating Values of Models

Several studies have evaluated LLMs on how they express social and cultural values in response to different prompts. A common approach involves using research instruments like Hofstede’s Value Survey Module (VSM) (Hofstede and Hofstede, 2016),

which has been applied in prior work (Kharchenko et al., 2024; Arora et al., 2023; Masoud et al., 2024) to assess whether models align their responses with cultural contexts. Despite differences in methodology, findings consistently show that LLMs adjust their value expressions based on contextual cues.

Other studies have constructed evaluation datasets based on the World Values Survey (Haerpfer et al., 2020), including GlobalOpinionQA (DURMUS et al., 2024) and WorldValueBench (Zhao et al., 2024). The former also incorporates value questions from Pew¹ finding that most LLMs tend to favor Western perspectives. The latter focuses on evaluating models’ awareness of demographic contexts, revealing that even advanced LLMs struggle to capture the nuances of multicultural value systems.

Recent studies have investigated model values within specific contexts. BiasLens (Li et al., 2024) systematically examines social biases in LLMs through role-playing scenarios, while Moore et al. (2024) evaluates value consistency across prompt variations, revealing generally stable model outputs. Liu et al. (2025) introduces a benchmark dataset to assess LLMs’ ability to infer implicit cultural values from natural conversational contexts, emphasizing the challenges of nuanced attitude detection and open-ended cultural reasoning.

Unlike prior studies, we analyze LLMs’ value expression patterns using multi-turn human–model interactions, which better reflect real-world chatbot inputs. We also evaluate behavioral consistency when sociodemographic attributes are provided explicitly versus implicitly within the dialogue.

3 Research Targets

We define *behavior adaptation capability* as a model’s ability to adjust response values and tone in accordance with users’ sociodemographic attributes. A key focus of our study is to examine whether this adaptation remains consistent when the same demographic information is supplied explicitly in a single-turn prompt versus implicitly via earlier dialogue (see Figure 1).

We use the Value Survey Module (VSM 2013) (Hofstede and Hofstede, 2016), grounded in Hofstede’s Cultural Dimensions Theory (Gerlach and Eriksson, 2021), to quantify cultural values. This questionnaire features multiple-choice items on workplace dynamics and decision-making, each

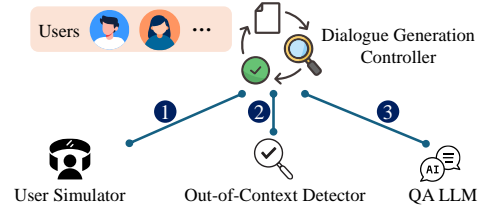


Figure 2: Dataset generation framework architecture. Each iteration: (i) user_simulator LLM is queried to generate a question simulating the user’s perspective based on their profile, (ii) out-of-context detector validates the question to ensure consistency with the user’s profile, and (iii) qa_llm responds to the question.

with five options (IDs 1–5). From the original 24 items, we select an 18-question subset Q , omitting emotional and health-related items. We apply this survey to evaluate model behavior adaptation across below experimental scenarios.

Behavior Adaptation to User Profile (BA_user):

This scenario evaluates whether models can adjust their responses based on explicit user profile consisting of sociodemographic attributes presented in the context, *e.g.*, “Answer questions based on the given user profile: age: 23, job title: data scientist, gender: male, education: Bachelor’s degree.”

Behavior Adaptation to Dialogue History (BA_dialogue):

Instead of relying on explicit attributes, models are tested on their ability to infer and adapt from dialogue history (Gupta et al., 2024). This scenario mimics real-world interactions, where the model must interpret user intent and context from prior exchanges.

Consistency Across Profile and Dialogue History (Consistency):

Beyond behavioral adaptation, our framework also evaluates whether models maintain behavioral consistency when processing equivalent user attributes presented in different representational formats. Specifically, we expect models to respond similarly to the value survey when the same demographic attributes are provided through explicit user profiles or implicitly inferred from dialogue history.

4 Dialogue Dataset Generation

While BA_user assessment is straightforward, evaluating BA_dialogue and Consistency is hindered by the absence of datasets meeting two key criteria: (1) human–model dialogues with organically embedded demographic attributes, and (2) explicit

¹<https://www.pewresearch.org/>

mappings between each dialogue and its corresponding user profile.

Drawing inspiration from prior works (Abdullin et al., 2024; Chen et al., 2024), we design a multi-agent workflow to generate synthetic, career-advice dialogues from user profiles sourced from a curated simulated dataset.² Each dialogue maps to a unique profile, embedding demographic attributes—age, education, occupation, and nationality—within contextually grounded interactions. This career-focused domain aligns with our value assessment framework, enhancing interpretability in downstream evaluations.

The dialogue is generated iteratively by the workflow under the supervision of a generation controller, as illustrated in Figure 2. The controller orchestrates three key LLM components:

User Simulator (user_simulator) We employ Gpt-4o-2024-08-06 (OpenAI et al., 2024) to emulate a user seeking career advice via question-answer (QA) interactions with an LLM. Each query generated by the simulator is guided by: (1) user demographic attributes for personalization, (2) instructions and preceding dialogue, guiding the generation of contextually relevant queries, and (3) predefined conversation objectives and termination criteria. The simulator ends the dialogue once the specified termination condition is satisfied.

Out-of-Context Detector (ooc_detector): We employ Gpt-4o-mini-2024-07-18 (OpenAI et al., 2024) to validate the questions generated by the user simulator. It ensures that each question aligns with the user’s profile and maintains consistent first-person framing. If inconsistencies are detected, the ooc_detector directly revises the question.

Question Answering LLM (qa_llm): The LLM responds to the simulated user’s queries with its default configuration to ensure natural interactions. To replicate real-world chatbot behavior, past dialogue history is always prepended to the latest user question, following standard practices for context injection in human-LLM communication.

Each generation loop terminates when the user simulator decides to conclude the dialogue or the maximum iteration limit (max_runs) is reached.

More details on the prompt design for each component are provided in Figure 6 (Appendix B), and the complete conversation generation procedure is outlined in Algorithm 1 (Appendix C). In total,

Dimensions	LLM Judge	Human
Attribute Coverage	4.14	3.64
Attribute Correctness	4.76	4.97
Question Diversity	4.52	4.65
Relevance	4.63	4.26

Table 1: Overall ratings for generated dialogues by the LLM judge and 50 human-rated samples. Despite stricter human judgments, high scores across all four dimensions confirm dialogue quality.

1000 dialogue sets, denoted by D , are generated,³ each mapped to a unique user from the seed dataset. The full user set is denoted by U .

4.1 Dataset Evaluation

To assess the quality of the generated dataset, we conduct both human and LLM-based evaluations. LLM evaluation follows the widely adopted “LLM-as-a-judge” methodology (Zheng et al., 2023; Gu et al., 2025). Only questions generated by the user_simulator are assessed, as the qa_llm role plays as itself to respond to these simulated queries. Assessments are conducted across four dimensions:

Attribute Coverage. The number of demographic attributes explicitly mentioned (up to 5).

Attribute Correctness. The number of correctly referenced demographic values. For example, if the user’s age, gender, and job title are mentioned, but an incorrect age is used, the score for Attribute Coverage is 3 and for Attribute Correctness is 2.

Question Diversity. The variety of topics covered, reflects the simulator’s ability to generate distinct, contextually rich questions. For example, if four questions are generated but all focus solely on short-term career advice, then the score is 1.

Relevance. The extent to which the questions remain contextually appropriate for career advice, and align with the qa_llm’s prior responses to maintain coherent conversational flow.

For human evaluation, three postgraduate annotators score the generated questions on a 0 to 5 scale, with higher values indicating better quality. All annotators independently assess the same subset of 50 randomly selected samples. For automated evaluation, we use gpt-4o-2024-08-06 as the judge to score all samples. Identical scoring guidelines are supplied to both human and LLM raters (Appendix D), in accordance with best practices from Leng (2023). To reduce variability, the

²<https://www.kaggle.com/datasets/ravindrasinghrana/employeeedataset/data>

³https://github.com/FerdinandZhong/model_behavior_adaption

LLM judge evaluates each sample with 10 different random seeds, and we report the average score.

We assess alignment between average human and LLM judge ratings on the shared subset using the Pearson correlation coefficient (Freedman et al., 2007) and the two-way mixed-effects intraclass correlation coefficient (ICC(3, k)) (Shrout and Fleiss, 1979). We omit Fleiss’ Kappa due to its sensitivity to category prevalence in our skewed data (Hoehler, 2000). Results (Appendix E) demonstrate strong concordance between human and automated evaluations. Summary statistics for all four evaluation dimensions are presented in Table 1, confirming the high quality of the generated dataset.

5 Behavior Adaptation Evaluation

Using our synthetic dataset, we assess LLM behavior adaptation via two scenarios.

In the BA_user scenario, each time a model is queried with a VSM question $q_j \in Q$ and an explicit user profile $u \in U$, where q_j denotes the j th question. The model must return a selected_option_id (1–5) corresponding to the question’s option_ids—the IDs of available choices—along with a justification and the log probability distribution over option_ids. We denote this composite response by r_u^j .

In the BA_dialogue scenario, each time a model is presented with a synthetic dialogue $d \in D$ followed by a VSM question q_j , it generates a response r_d^j under the same requirements. The complete sets of responses $\{r_u^j\}$ and $\{r_d^j\}$ for each model are denoted by R_U and R_D , respectively. The querying workflow is illustrated in Figure 3.

5.1 Distance Definition

The core component of each response r is a normalized probability distribution P over all possible option_ids. This distribution is computed from the model’s log probability outputs (see Appendix G for details). For instance, if a response r has a selected_option_id of 2, the distribution might be $[0.1, 0.7, 0.05, 0.0, 0.15]$, where each value represents the relative likelihood of the corresponding option_id. After normalization, the total probability sums up to 1.

We quantify the unit divergence between two responses r ’s by calculating the Jensen-Shannon divergence (JSD) between their corresponding probability distribution P ’s. That is, $\text{JSD}(P_u^j \parallel P_{u'}^j)$ donates the divergence between r_u^j and $r_{u'}^j$, where

u and u' refer to two distinct user profiles.

To assess BA_user and BA_dialogue, both U and D are partitioned into groups g based on attributes including age, education, occupation, and nationality. For each group and question index j , we compute the Jensen-Shannon centroid c^j —the distribution minimizing the total JSD to all group responses P_i^j (Nielsen, 2020). Starting from the mean distribution \bar{P}^j , the centroid is obtained by

$$c^j = \arg \min_c \sum_{i=1}^n \text{JSD}(c \parallel P_i^j),$$

where P_i^j is the i th response distribution for question j in a group of size n .

The overall divergence between two groups g and g' is then defined as:

$$\text{Distance}(g, g') = \frac{1}{|J|} \sum_{j=1}^{|J|} \text{JSD}(c_g^j \parallel c_{g'}^j),$$

This centroid-based approach helps mitigate the effects of outliers and uneven group sizes.

We establish baseline values for both BA_user and BA_dialogue using a consistent methodology. To illustrate, consider the baseline for BA_user evaluation. After grouping R_U by sociodemographic attribute and computing a centroid for each group, we define the global centroid as the Jensen-Shannon centroid of all responses in R_U . The baseline is then calculated as the average distance between each group centroid and the global centroid. Comparing this baseline to the inter-group divergences allows us to quantify the model’s sensitivity to demographic variation. For example, when grouping R_U by age, the divergence between the “<30” and “>60” cohorts should noticeably exceed the baseline, while the divergence between “<30” and “30–40” should remain below it—illustrating that greater age gaps drive greater variation in generated responses.

To evaluate the Consistency scenario, we compare responses R_u and R_d corresponding to the same user, by using their selected_option_id sequences rather than the full probability distributions. Let $S_u = [s_1, s_2, \dots, s_m]$, and $S_d = [t_1, t_2, \dots, t_m]$, where s_j is the selected_option_id in 1 to 5 for the j th question (the same applies to t_j), and m is the total number of questions, and $m = 18$ in our setting.

Since the option_ids are ordinal and questions are independent, we employ the Earth Mover’s Dis-

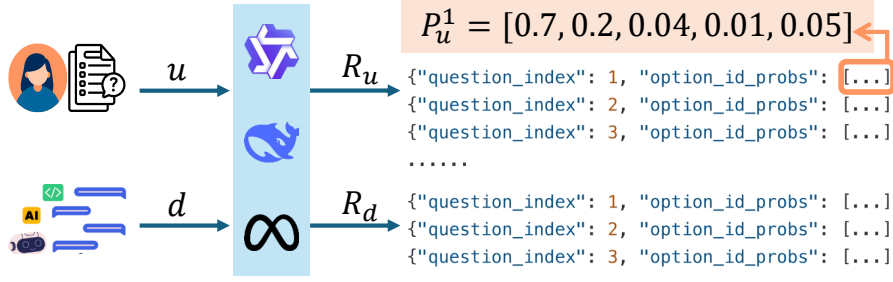


Figure 3: Model querying workflow with key components. Here, u denotes a user profile, and d is a synthetic dialogue. Each line $r \in R$ represents the response to a VSM question, which includes a normalized probability distribution P over the 5 option_ids.

tance (EMD) (Rubner et al., 1998) to quantify alignment between S_u and S_d . Let $h_u(k)$ and $h_d(k)$ denote the frequencies of value k in S_u and S_d , respectively, and let $K = 5$ be the total number of discrete bins. Then

$$\text{EMD}(S_u, S_d) = \sum_{x=1}^{K-1} \left| \sum_{k=1}^x \frac{h_u(k)}{m} - \sum_{k=1}^x \frac{h_d(k)}{m} \right|.$$

We define a model’s Consistency as the mean EMD across all matched user profile u and his/her dialogue d pairs or (u, d) ’s, with a baseline computed over an equivalent number of random (u, d) pairings *i.e.*, randomly matching a user profile with another user’s dialogue. This metric enables direct comparison of consistency when the same profile is presented in different formats.

Table 2 provides the details of the measured divergences and baseline values for all scenarios.

5.2 Experiment Setting

Following the workflow outlined in Figure 3, we evaluate multiple open-source LLMs, including the *Qwen2.5* family, *Llama3.1* family, *DeepSeek-V3*, and the reasoning model *QwQ-32B* (Qwen et al., 2025; Dubey et al., 2024; DeepSeek-AI, 2024), to generate R_U and R_D . Model outputs are structured in JSON format using XGrammar (Dong et al., 2024) for streamlined processing. Each model is queried once per prompt, except for the reasoning model *QwQ-32B*, which is queried twice. The first query is unconstrained to allow for reasoning content generation, which is then appended to the second query to elicit a structured response. Full prompt designs are listed in Appendix F.

6 Evaluation Outcomes

Recall that the selected ID is the one assigned the highest probability by the model. The probability

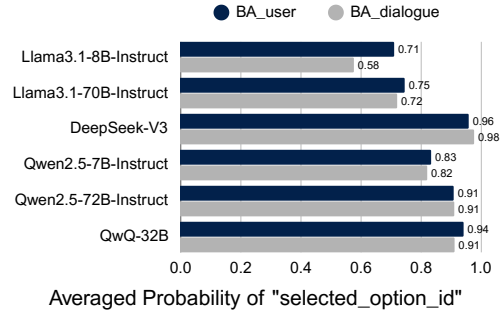


Figure 4: Mean probability of the “selected_option_id” in BA_user and BA_dialogue, reflecting model confidence. Most models show similar decisiveness across both scenarios.

value associated with this ID reflects the model’s confidence in its selection. We treat this probability as the confidence score for the selected ID and use the average of these scores across all responses to estimate the model’s overall confidence. As illustrated in Figure 4, all models—except *Llama3.1-8B-Instruct*, which shows slightly reduced confidence in the dialogue setting—exhibit consistently high confidence across both contexts, supporting the interpretation that their selections reflect genuine preferences and reinforcing the validity of our subsequent analyses.

6.1 Behavior Adaptation

Next, we assess behavior adaptation by varying demographic attributes. Responses in R_U and R_D are grouped by age, education, occupation, and nationality, and we compute inter-group divergences and corresponding baselines as specified in Table 2. Adaptation is quantified as the ratio of each divergence to its baseline (baseline=1). Detailed analyses for age and education appear below; further results are available in Appendix H.

Age: We group responses R_U and R_D into 10-

Scenarios	Measured Distance	Baseline
BA_user	$\frac{1}{ J } \sum_{j=1}^{ J } \text{JSD}(c_g^j \parallel c_{g'}^j), \quad g, g' \subseteq U$	$\frac{1}{ J } \frac{1}{ G } \sum_{j=1}^{ J } \sum_{g=1}^{ G } \text{JSD}(c_g^j \parallel c_U^j)$
BA_dialogue	$\frac{1}{ J } \sum_{j=1}^{ J } \text{JSD}(c_g^j \parallel c_{g'}^j), \quad g, g' \subseteq D$	$\frac{1}{ J } \frac{1}{ G } \sum_{j=1}^{ J } \sum_{g=1}^{ G } \text{JSD}(c_g^j \parallel c_D^j)$
Consistency	$\frac{1}{ U } \sum_{i=1}^{ U } \text{EMD}(S_u, S_d)$	$\frac{1}{ U } \sum_{i=1}^{ U } \mathbb{I}(u_i \not\sim d_i) \text{EMD}(S_u, S_d)$

Table 2: List of quantified divergences and baselines across evaluation scenarios. For BA_user and BA_dialogue, responses are partitioned into attribute-based groups g (**age**, **education level**, **occupation**, and **country**) to assess their influence on model behavior. Scenario Consistency measures the alignment between user profiles u and their corresponding dialogue histories d . J represents the set of questions. $u_i \not\sim d_i$ means the user profile u_i and dialogue d_i are not from a same user.

Models	Measured Distance ↓	Baseline	Distance/Baseline ↓
Llama3.1-8B-Instruct	0.305	0.312	0.978
Llama3.1-70B-Instruct	0.214	0.225	0.951
Qwen2.5-7B-Instruct	0.276	0.276	1.000
Qwen2.5-72B-Instruct	0.176	0.190	0.926
DeepSeek-V3	0.118	0.128	0.922
QwQ-32B	0.112	0.125	0.896

Table 3: Results of evaluating LLMs’ capability to maintain response consistency across different context formats. Both absolute distances and relative divergence ratios over the baseline are displayed in the table. QwQ-32B has both the lowest absolute distance values and the lowest relative value of distance over baseline.

year brackets (*e.g.*, “30–40”), with “< 30” and “> 60” as boundary categories. Figure 5 (first row) presents results for both BA_user and BA_dialogue. In both scenarios, models exhibit a positive correlation between age disparity and behavioral divergence (*e.g.*, maximum divergence between “< 30” and “> 60”). Notably, models exhibit greater consistency in the BA_dialogue scenario than in BA_user, indicating that age information conveyed through dialogue enables more stable adaptation across age groups.

Education Level: Educational attainment provides a precise grouping criterion. Figure 5 (second row) reports the divergences and baselines for both BA_user and BA_dialogue. As with age, models exhibit greater divergence in responses’ values as educational disparity increases (*e.g.*, highest divergence between high school and doctoral degrees), whether the attribute is provided explicitly or via dialogue. These results further demonstrate models’ capacity to incorporate educational background into value and nuance adaptation.

Additional evaluations using other attribute-based groupings are presented in Appendix H. These findings show that most LLMs accurately infer sociodemographic attributes and adjust their responses accordingly, with larger attribute differences producing greater shifts in expressed values. Crucially, despite dialogue history being less ex-

plicit than user profiles, models still adapt their behavior to align with user characteristics.

6.2 Consistency across Context Formats

We next evaluate each model’s behavioral consistency across context formats (scenario Consistency). Following Section 5.1, we compute, for each model, both the distance and its baseline. While absolute distances quantify response variability across formats, they may conflate behavioral alignment with format-induced noise (He et al., 2024). To disentangle these effects, we also report the ratio of measured distance to baseline: lower ratios signify stronger consistency in adapting to demographic cues across differing formats.

Table 3 summarizes these metrics. Smaller tested models (with parameters smaller than 10B) exhibit both higher absolute EMDs and ratios near or equal to 1, indicating that prompt-format variability outweighs demographic alignment. In contrast, larger models achieve lower ratios, reflecting greater robustness to format changes and improved consistency in behavior adaptation. Furthermore, we observe a positive correlation between the measured consistency and benchmark performance on language understanding and reasoning tasks (*e.g.*, MMLU Pro (Wang et al., 2024), Big-Bench Hard (Suzgun et al., 2022)): models with superior cognitive capabilities tend to maintain higher

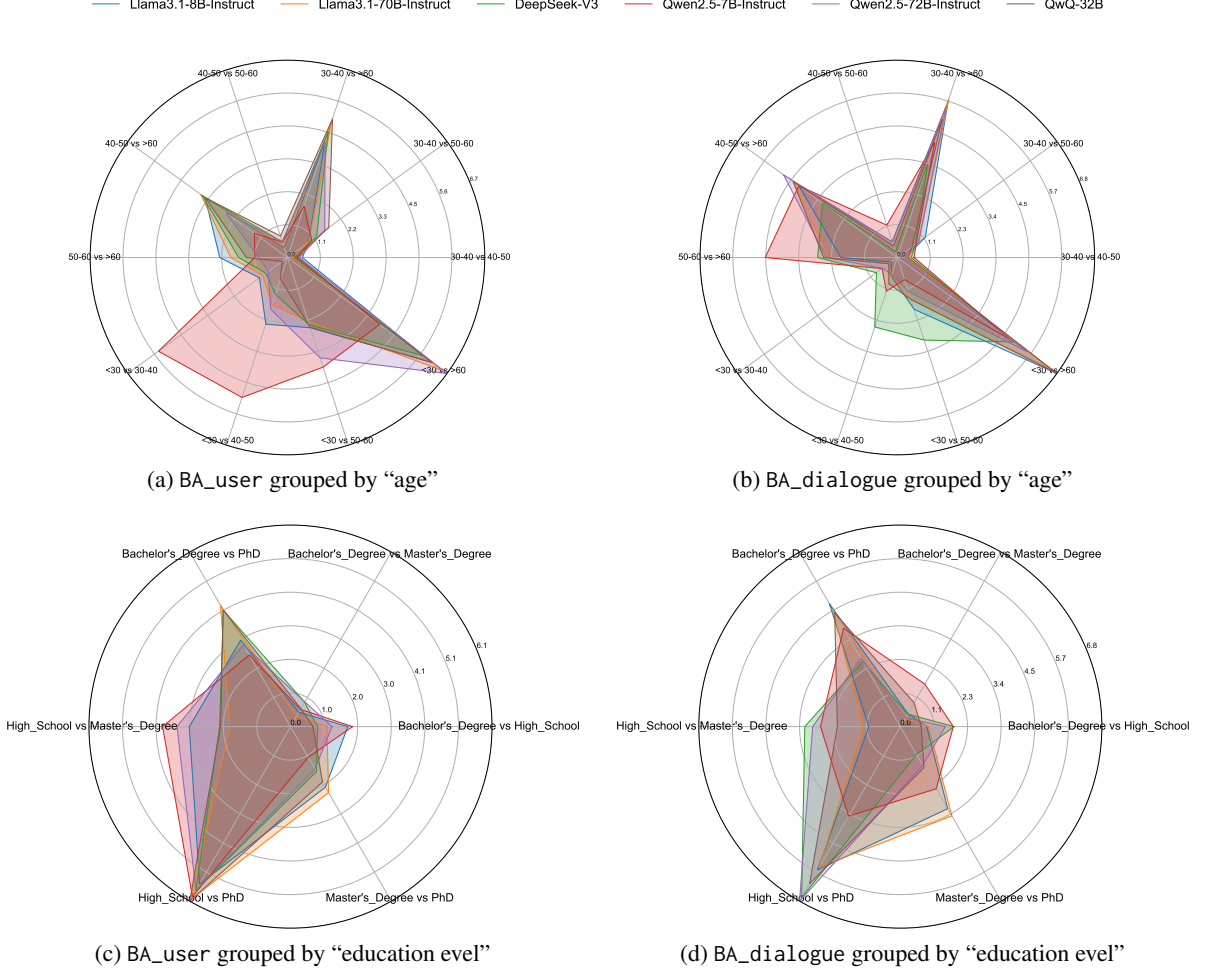


Figure 5: The measurement results for BA_user and BA_dialogue are shown below. The first row compares groups by "Age," while the second row presents results for "Education Level." Most models exhibit a positive correlation between computed distances and demographic differences.

consistency across formats.

The reasoning-augmented model *QwQ-32B*, specifically trained through reinforcement learning for reasoning, achieves the highest consistency, despite its relatively smaller size. We attribute this superior consistency to its enhanced reasoning capabilities. To substantiate this, we recorded and analyzed its reasoning traces, revealing that *QwQ-32B* systematically revisits and integrates the provided sociodemographic attributes, whether explicitly stated or inferred from dialogue history. This iterative process enables the model to select responses more precisely aligned with the given attributes, promoting consistency across context formats. The ability to methodically revisit and incorporate attribute details likely contributes to its superior alignment performance. Reasoning process samples are provided in Appendix I.

7 Conclusion

This study presents a novel framework for evaluating how LLMs adapt their behavioral outputs when sociodemographic attributes are provided through two distinct interaction formats: (i) explicit single-turn prompts and (ii) multi-turn dialogue integration. We then systematically measure cross-format consistency to assess robust sociodemographic adaptation. Our findings indicate that most models adjust effectively to single-format attribute changes, particularly in attributes like age and education level, with the degree of value adjustment positively correlated with the magnitude of attribute change. However, significant discrepancies arise in cross-format scenarios. Smaller models often struggle to maintain consistent alignment across formats, while larger, reasoning-augmented models demonstrate more stable performance.

Limitations

This study has a few limitations that require further investigation in future research.

- **Prompt Design:** Prompts for querying models to determine the most suitable response are designed based on our experience and experiments with sample cases, given the lack of established guidelines for optimizing prompts in value survey question answering. Future research could investigate how prompt content variations (besides the embedded information) impact models' behavior adaptation.
- **Limited Scope of Value Survey:** The value survey used to assess model behavior consists of a relatively small number of questions from the VSM 2013 survey, which has been criticized for its simplicity and limited scope. However, it aligns well with our generated dataset, as it focuses on values reflected in career-related questions. Future studies could enhance the evaluation by incorporating larger and more diverse question sets.
- **Single Source of Dialogue:** The evaluation of BA_dialogue is based on dialogues generated by GPT-4O as the QA bot. When providing the tested models with the dialogue history, GPT's specific response styles may influence the behavior of the tested models, which could be further explored in future work.

Ethical Considerations

Our study examines how LLMs adapt outputs to users' sociodemographic contexts through explicit (profile-based) and implicit (dialogue-embedded) input formats. Misalignment or inconsistencies in model outputs can reinforce stereotypes or erode user trust in cross-cultural interactions. To enable controlled, privacy-preserving evaluations, we curate and open-source a synthetic dataset, with all user profiles derived from a publicly available synthetic dataset (see Section 4), thereby eliminating any risk of personal identifying information (PII) exposure or consent violations. However, given the dataset's controlled topical focus, excessive reliance on these examples without careful consideration of their suitability for use cases may introduce bias and overlook the broader diversity of real-world interactions.

References

- Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. [Synthetic dialogue dataset generation using llm agents](#). *Preprint*, arXiv:2401.17461.
- Kristen M. Altenburger, Hongda Jiang, Robert E. Kraut, Yi-Chia Wang, and Jane Dwivedi-Yu. 2024. [Examining the role of relationship alignment in large language models](#). *Preprint*, arXiv:2410.01708.
- Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2023. [Probing pre-trained language models for cross-cultural differences in values](#). *Preprint*, arXiv:2203.13722.
- Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. 2024. [DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9057–9079, Miami, Florida, USA. Association for Computational Linguistics.
- Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. [A complete survey on llm-based ai chatbots](#). *Preprint*, arXiv:2406.16937.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. [Xgrammar: Flexible and efficient structured generation engine for large language models](#). *Preprint*, arXiv:2411.15100.
- Abhimanyu Dubey and Abhinav Jauhri et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askill, 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. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling*.
- David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics (international student edition)*. *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.
- Helene H Fung, Yuan Wan Ho, Rui Zhang, Xin Zhang, Kimberly A Noels, and Kim-Pong Tam. 2016. Age differences in personal values: Universal or cultural specific? *Psychol. Aging*, 31(3):274–286.
- Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. [LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.
- Michele J Gelfand and Jana L et al. Raver. 2011. Differences between tight and loose cultures: a 33-nation study. *Science*, 332(6033):1100–1104.
- Philipp Gerlach and Kimmo Eriksson. 2021. Measuring cultural dimensions: External validity and internal consistency of hofstede’s VSM 2013 scales. *Front. Psychol.*, 12:662604.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. *A survey on llm-as-a-judge*. *Preprint*, arXiv:2411.15594.
- Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. 2024. *LLM task interference: An initial study on the impact of task-switch in conversational history*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14633–14652, Miami, Florida, USA. Association for Computational Linguistics.
- C Haerpfer, R Inglehart, A Moreno, C Welzel, K Kizilova, J Diez-Medrano, M Lagos, and P Norris. 2020. *World Values Survey: Round Seven - Country-Pooled Datafile*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. *Does prompt formatting have any impact on llm performance?* *Preprint*, arXiv:2411.10541.
- Fred K. Hoehler. 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of clinical epidemiology*, 53 5:499–503.
- G Hofstede and G. J. Hofstede. 2016. VSM 2013. <https://geerthofstede.com/research-and-vsm/vsm-2013/>. Accessed: 2024-1-11.
- Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. *Faithful persona-based conversational dataset generation with large language models*. *Preprint*, arXiv:2312.10007.
- Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. *How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions*. *Preprint*, arXiv:2406.14805.
- Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. *Large language models as superpositions of cultural perspectives*. *Preprint*, arXiv:2307.07870.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Quinn Leng. 2023. Best practices for LLM evaluation of RAG applications: A case study on the databricks documentation bot. <https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG>. Accessed: 2025-04-13.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. *BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. *CoRR*, abs/1910.13461.
- Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Yiling Lou, Tianlin Li, Weisong Sun, Yang Liu, and Xuanzhe Liu. 2024. *Benchmarking bias in large language models during role-playing*. *Preprint*, arXiv:2411.00585.
- Ziyi Liu, Priyanka Dey, Zhenyu Zhao, Jen tse Huang, Rahul Gupta, Yang Liu, and Jieyu Zhao. 2025. *Can llms grasp implicit cultural values? benchmarking llms’ metacognitive cultural intelligence with cq-bench*. *Preprint*, arXiv:2504.01127.
- Vera Lomazzi and Daniel Seddig. 2020. *Gender role attitudes in the international social survey programme: Cross-national comparability and relationships to cultural values*. *Cross-Cultural Research*, 54:106939712091545.
- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. *Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions*. *Preprint*, arXiv:2309.12342.
- Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. *Are large language models consistent over value-laden questions?* In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.
- Frank Nielsen. 2020. *On a generalization of the jensen–shannon divergence and the jensen–shannon centroid*. *Entropy*, 22(2).
- OpenAI, Josh Achiam, and Steven Adler et al. 2024. *Gpt-4 technical report*. *Preprint*, arXiv:2303.08774.
- Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. *Pchatbot: A large-scale dataset for personalized chatbot*. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 2470–2477, New York, NY, USA. Association for Computing Machinery.

- Qwen, :, and An Yang et al. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Y. Rubner, C. Tomasi, and L.J. Guibas. 1998. [A metric for distributions with applications to image databases](#). In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66.
- P E Shrout and J L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.*, 86(2):420–428.
- Anthony Sicilia, Jennifer C. Gates, and Malihe Alikhani. 2024. [Humbel: A human-in-the-loop approach for evaluating demographic factors of language models in human-machine conversations](#). *Preprint*, arXiv:2305.14195.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. [An evaluation of cultural value alignment in llm](#). *Preprint*, arXiv:2504.08863.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *Preprint*, arXiv:2210.09261.
- Jean M. Twenge. 2017. Have smartphones destroyed a generation? *The Atlantic*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhramil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Preprint*, arXiv:2406.01574.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Huggingface’s transformers: State-of-the-art natural language processing](#). *Preprint*, arXiv:1910.03771.
- 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.
- Jing Yao, Xiaoyuan Yi, and Xing Xie. 2024. [Clave: An adaptive framework for evaluating values of llm generated responses](#). *Preprint*, arXiv:2407.10725.
- 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.
- Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. 2024. [Heterogeneous value alignment evaluation for large language models](#). *Preprint*, arXiv:2305.17147.
- Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. [World-valuesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models](#). *Preprint*, arXiv:2404.16308.
- 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.

A Capability of LLMs for Recognizing Persona Attributes

As a preliminary stage of this study, we select Llama3.1-8B-Instruct as a representative model to verify its ability to extract user attributes from dialogue. This model is one of the smallest evaluated in our study. For evaluation, we utilize the FaithfulPersonaChat (Jandaghi et al., 2023) synthetic dataset, which comprises 5,648 unique synthetic personas and 11,001 conversations. Each conversation takes place between two users (User1 and User2), and the persona profiles of both participants are provided alongside the dialogue. Every synthetic user is assigned five persona attributes, with at least one of these attributes explicitly mentioned in their dialogues.

Given the dataset structure, we prompt the tested model with a conversation and ask it to identify one mentioned persona attribute for each user (evaluated one at a time). The model selects from a list of ten candidate attributes, which includes the user’s five correct attributes and five randomly sampled attributes from other users. The selected attribute is then evaluated against two reference settings: (1) the ground truth, i.e., the user’s complete set of five correct attributes, and (2) the output from gpt-4o-mini (OpenAI et al., 2024), which is prompted to extract all identifiable persona attributes from the conversation. If the selected attribute appears in either reference, the model receives a score of 1; otherwise, the score is 0. The final score is calculated by aggregating the hit rate across all 11,001 conversations for each user. We test the model under different prompting strategies, ranging from zero-shot to ten-shot settings.

The results, shown in Table 4, confirm that the model can *effectively identify at least one persona attribute* from dialogues. This validation supports our main research objective: to investigate whether models can adapt their responses based on the recognized attributes.

B Prompts Design for Dataset Generation Agent

User Simulator (user_simulator): We design two types of prompts—initial and subsequent—as illustrated in Figure 6 to guide the LLM in simulating the user for question generation. The initial prompt starts the conversation, while subsequent prompts incorporate chat history to maintain context. This iterative approach enables the

generation of relevant follow-up questions and ensures dialogue progression. To prevent the user_simulator from misinterpreting its role, the generated dialogue is embedded within a single message block in subsequent prompts.

Question Answering LLM (qa_llm): No additional prompt design is applied when querying the LLM in the role of a QA agent. It is treated as a standard chatbot, receiving the question generated by the user_simulator directly.

Out-of-Context Detector (ooc_detector): We specify two criteria for evaluating the generated question in the system prompt for ooc_detector: (1) whether the question accurately reflects the user profile information and (2) whether it is framed in the first person. The user profile and generated question are provided in the user prompt. This design is illustrated in Figure 7.

C Pseudocode for Dataset Generation

The complete procedure for generating the dialogue dataset is outlined in the pseudocode presented in Algorithm 1. During dialogue generation, we set the threshold *dialogue_max_runs* to 5.

D Prompts Design for LLM Judge

To enhance the reliability of the LLM judge when evaluating questions generated by user_simulator, we provide detailed scoring instructions in the prompt, including explicit criteria for each possible score. Additionally, we require the model to articulate its reasoning behind each score, encouraging a more deliberate evaluation process. Prompts of all 4 dimensions are listed in Figures 8, 9, 10, 11

E Alignment between LLM Judge and Human Rates

We measure the alignment between the scores assigned by the LLM judge and the average human ratings across 50 sampled dialogues. The results, presented in Table 5, show strong alignment for Dimensions 1, 3, and 4. The relatively lower alignment for *Attribute Correctness* is likely due to high overall scores and low variance in this dimension, as the correctness of attributes in the generated dialogues is largely ensured by the robust user_simulator and additional validation performed by the out-of-context detector in the dialogue generation pipeline.

Prompt Type	Models	Hit Rate			
		vs Ground Truth		vs gpt-4o-mini-2024-07-18	
		User1	User2	User1	User2
Zero-shot	Llama3-8B-Instruct	0.826	0.864	0.732	0.784
	Llama3.1-8B-Instruct	0.821	0.864	0.731	0.796
One-shot	Llama3-8B-Instruct	0.849	0.895	0.775	0.839
	Llama3.1-8B-Instruct	0.837	0.888	0.756	0.827
Five-shot	Llama3-8B-Instruct	0.851	0.906	0.784	0.854
	Llama3.1-8B-Instruct	0.848	0.905	0.774	0.853
Ten-shot	Llama3-8B-Instruct	0.856	0.910	0.783	0.855
	Llama3.1-8B-Instruct	0.852	0.905	0.775	0.853
Random	–	0.5	0.5	0.18	0.19

Table 4: Recognition accuracy of persona attributes from dialogue by LLMs. We evaluate both “Llama3.1-8B-Instruct” and “Llama3-8B-Instruct”, which belong to the same Llama model family. Model scores are compared against the expected accuracy of randomly selecting one correct attribute from the list of 10 candidates.

Procedure 1 Generation of Dialogues

Require: *seed_dataset*, *user_simulator*, *qa_llm*, *ooc_detector*, *dialogue_max_runs*

```

1: initialize dialogues
2: for each user_profile in seed_dataset do
3:   initialize conversation_history
4:   user_question ← user_simulator(initial_prompt ∪ user_profile)
5:   if ooc_detector(user_question, user_profile) then
6:     continue                                ▷ Skip this profile if out-of-context (OOC)
7:   end if
8:   conversation_history.append(user_question)
9:   llm_output ← qa_llm(user_question)
10:  conversation_history.append(llm_output)
11:  while True do
12:    user_question, end_conversation ← user_simulator(conversation_history ∪
    following_prompt ∪ user_profile)
13:    if end_conversation or len(conversation_history) ≥ dialogue_max_runs then
14:      dialogues.append(conversation_history)
15:      break                                ▷ End conversation or exceed max runs
16:    end if
17:    if ooc_detector(user_question, user_profile) then
18:      break                                ▷ Skip this profile if OOC detected
19:    end if
20:    conversation_history.append(user_question)
21:    llm_output ← qa_llm(user_question)
22:    conversation_history.append(llm_output)
23:  end while
24: end for

```

Dimensions	ICC(3, k)	Pearson Correlation
Attribute Coverage	0.85	0.75
Attribute Correctness	0.16	0.19
Question Diversity	0.63	0.55
Relevance	0.80	0.67

Table 5: Alignment scores, measured using ICC(3, k) and the Pearson Correlation Coefficient, were computed between the LLM judge and human annotators across four evaluation dimensions, based on a sample of 50 dialogues.

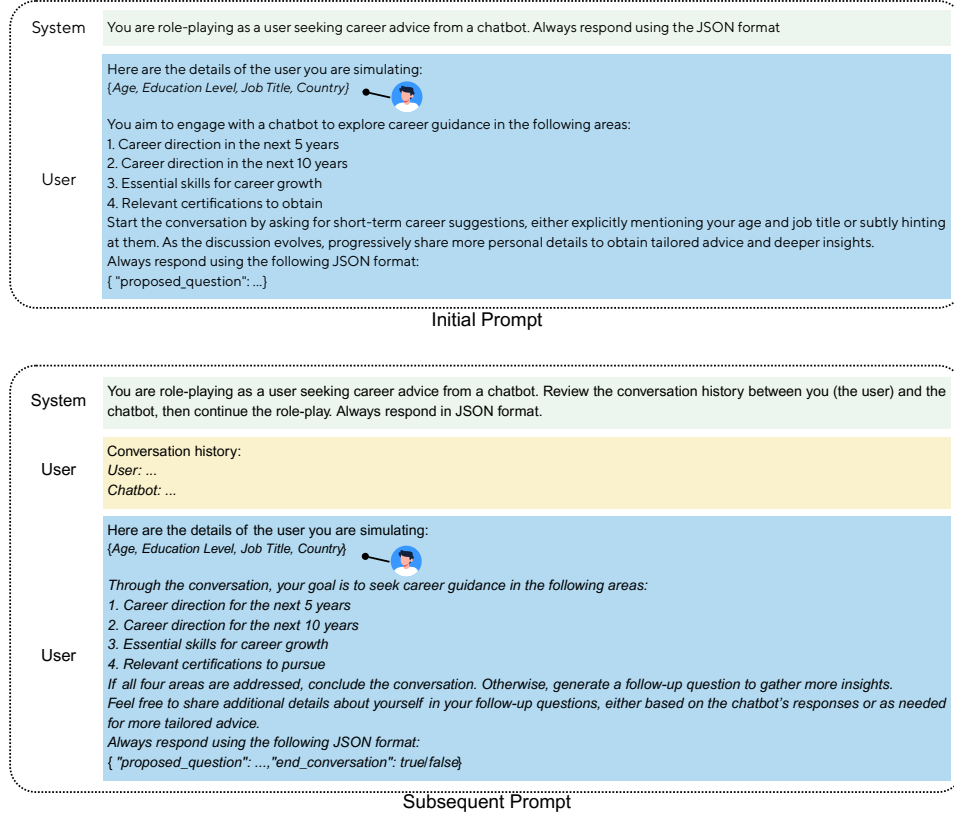


Figure 6: The prompts guiding the user_simulator are structured as follows: the **green text** is the system prompt defining the LLM’s role; the **blue text** specifies the user profile, conversation objectives, termination conditions, and response format; the **yellow text** depicts the chat history between the simulated user and the QA LLM.

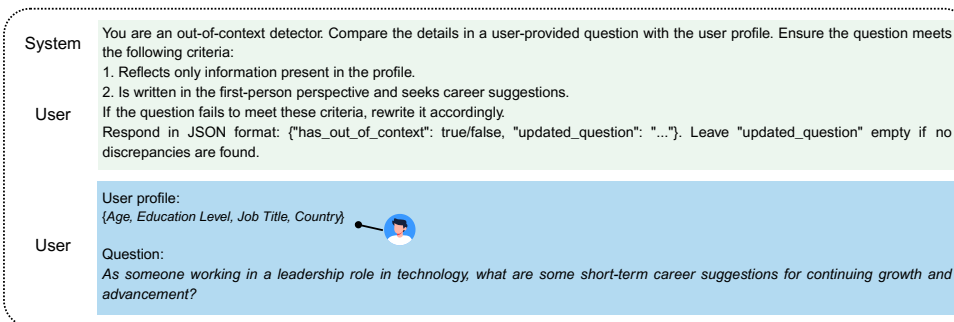


Figure 7: The prompt guiding the ooc_detector is structured as follows: the **green text** represents the system prompt, which defines the criteria for the LLM to check, while the **blue text** contains the user profile and the generated question.

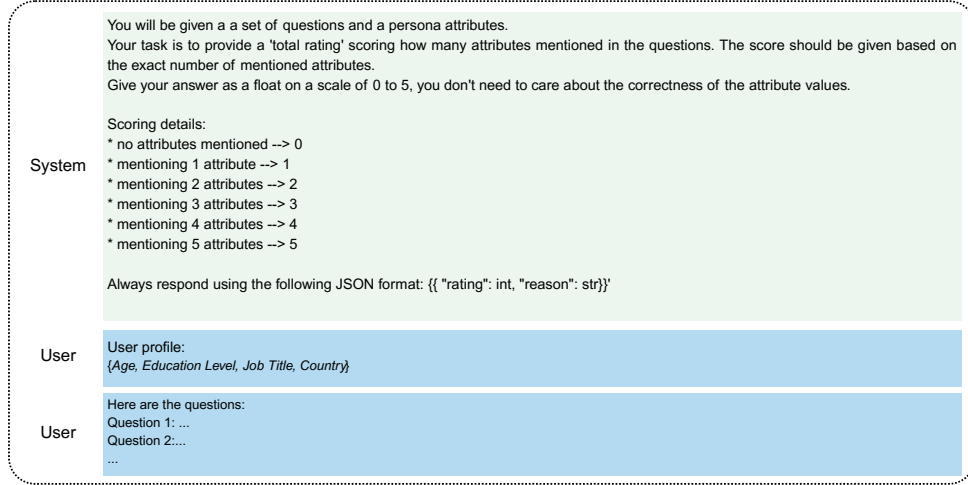


Figure 8: The prompt designed for guiding the LLM judge to score the generated dialogues from the perspective of “Attribute Coverage”

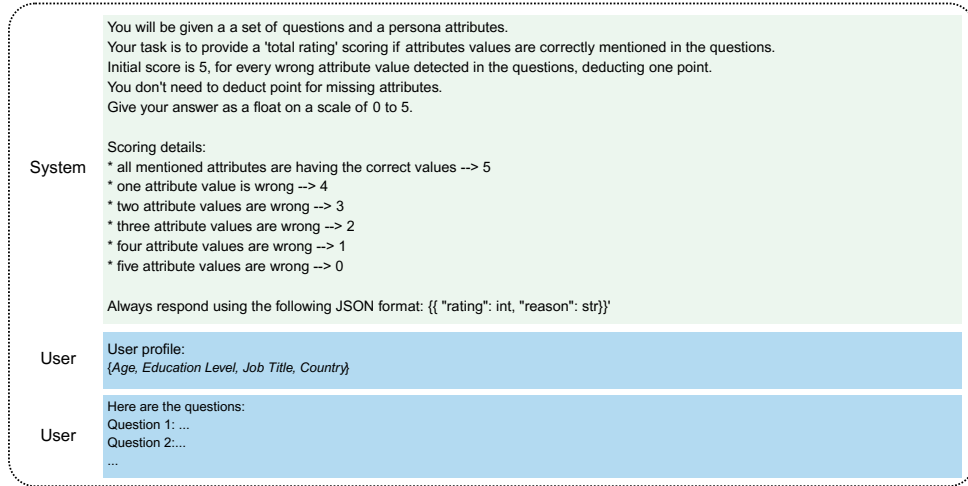


Figure 9: The prompt designed for guiding the LLM judge to score the generated dialogues from the perspective of “Attribute Correctness”

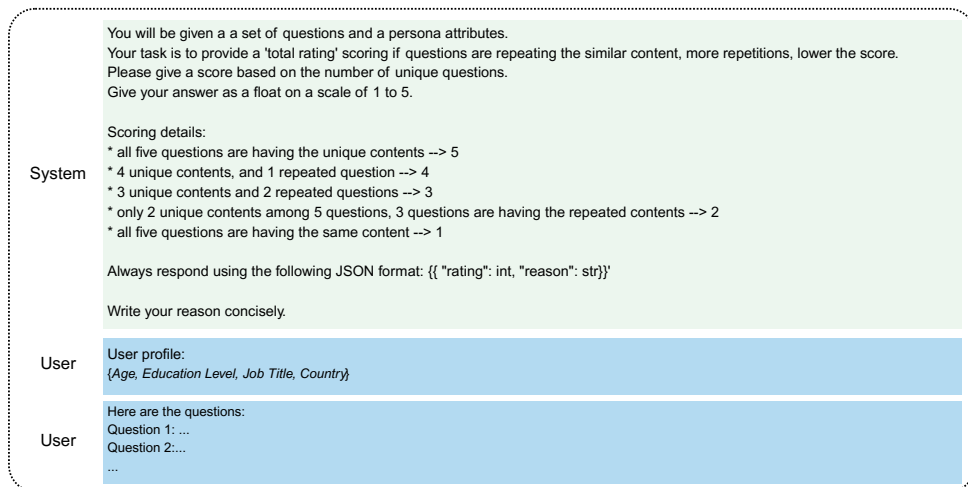


Figure 10: The prompt designed for guiding the LLM judge to score the generated dialogues from the perspective of “Question Diversity”

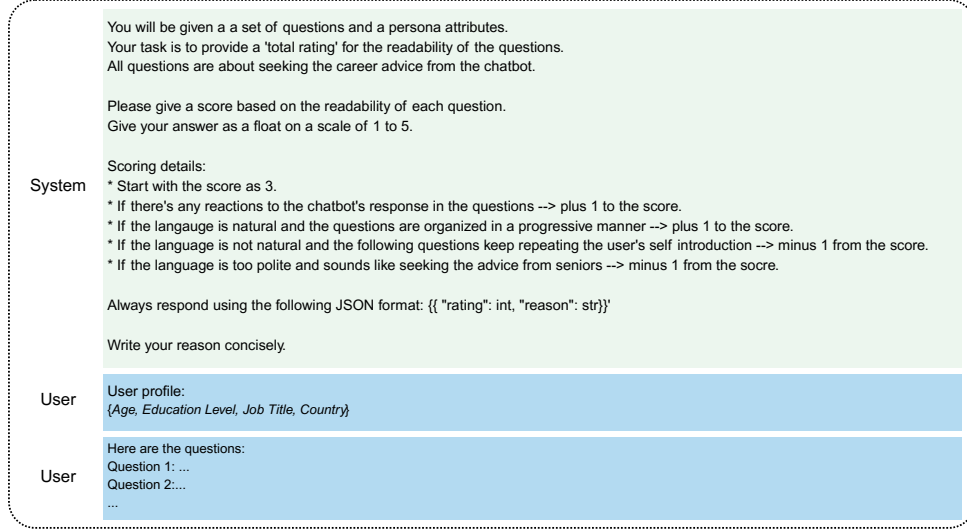


Figure 11: The prompt designed for guiding the LLM judge to score the generated dialogues from the perspective of “Relevance”

F Prompts Design for Querying Models

To ensure prompt naturalness, we devised two scenario-specific templates that vary in system prompt framing, information presentation, and question phrasing. The BA_user template directs the model: “*You are a chatbot designed to provide precise, personalized answers based on the given user profile,*” whereas the BA_dialogue template frames the model as the user: “*Assume you are me; help me find the most suitable answer to the following question.*” Full examples of both templates are provided in Figure 12. As described, reasoning models are queried twice per question to ensure a structured reasoning process. All prompts are crafted to naturally guide models in understanding embedded user attributes and making selections accordingly. To maintain coherence, questions containing explicit user attributes are presented in the third-person perspective, while those based on dialogue history adopt the first-person perspective, aligning with the conversational format.

G Probability Normalization for option_ids

The probability distribution P over option_ids for each response r is derived from the model’s log probabilities by applying the exponentiation function. Given that only the top 5 log probabilities from the full vocabulary are output by the model, any option_ids not included in this set are assigned a probability of 0.0. The resulting values are then normalized by dividing each probability

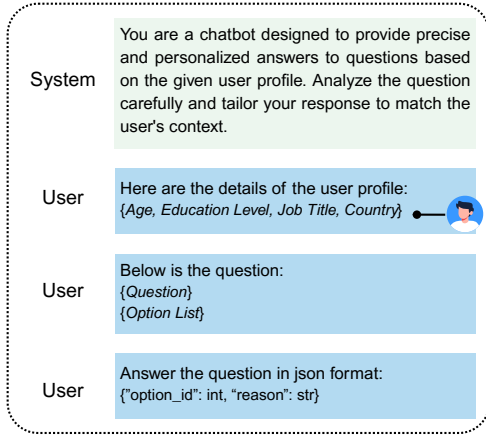
by the sum of all five option_ids probabilities, ensuring that the final distribution satisfies the condition:

$$\sum_{i=1}^5 p(i) = 1.$$

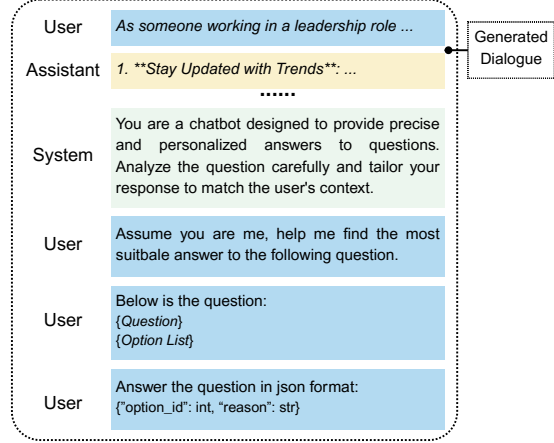
H Extra Evaluation Outcomes for BA_user and BA_dialogue

Nationality: Unlike attributes such as “Age” and “Education Level,” “Nationality” encompasses a large number of unique values. To address this, we categorize R_U and R_D based on each country’s “Development Level,” which we argue effectively captures cultural and value differences when incorporating country information into the model. The grouping is derived from a mapping list generated by the “gpt-4o-2024-08-06”. All tested models exhibit a strong awareness of country-based differences when classified into “Developed,” “Developing,” and “Third World” categories. Figure 13 visualizes this evaluation alongside analyses of alternative country-grouping methods.

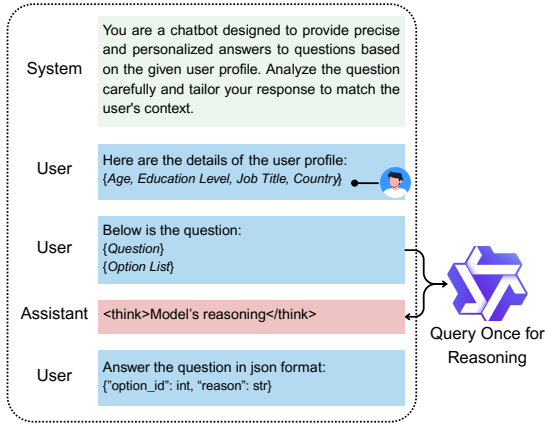
Job Category: The original dataset contains hundreds of distinct job titles. To facilitate analysis, we apply zero-shot classification using bart-large-mnli (Lewis et al., 2019), trained on MultiNLI (Williams et al., 2018), to map each title to a predefined set of job categories. We then group responses by these categories, as illustrated in Figure 14. While this approach introduces an



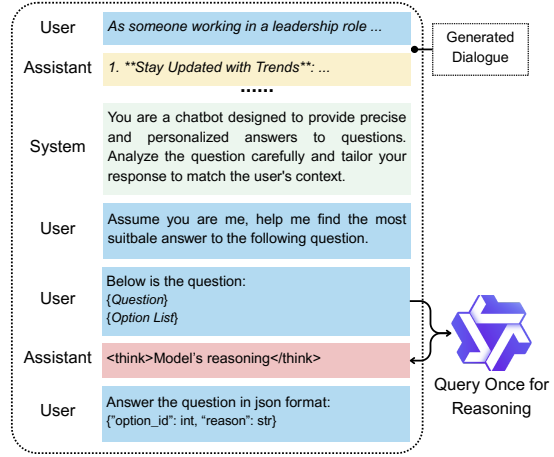
(a) Prompt Design for BA_user



(b) Prompt Design for BA_dialogue



(c) Prompt Design for BA_user



(d) Samples for BA_dialogue grouped by Education

Figure 12: Prompt designs for querying models to answer value-based questions given contextual information. Reasoning models are queried twice: the first query allows free-form reasoning responses, while the second query, with the reasoning content appended, enforces a structured output format.

additional layer of abstraction—potentially reducing the sensitivity to inter-group differences compared to attributes like age or education level—a consistent pattern emerges: responses to individuals in Science and Technology roles exhibit the greatest response similarity across both BA_user and BA_dialogue scenarios for most models. Conversely, when models identify users from Business-related sectors, the human values reflected in their responses tend to diverge more significantly from those associated with other sectors.

Position Level: Following a similar approach to grouping responses by job categories, we use `bart-large-mnli` to map job titles to predefined position levels, organizing the results accordingly. The outcomes are presented in Figure 15. Compared to the “Job Category” grouping, the position level analysis shows a more consistent pattern

across BA_user and BA_dialogue for all models. Notably, responses for “Entry Level” and “Senior Management” users exhibit the most pronounced differences, while the distances between “Entry Level” and “C-Suite” users are smaller than expected. This is likely because the classifier tends to assign only titles like “Chief of Staff” to the “Senior Management” category, while the “C-Suite” group contains a broader range of titles, resulting in a more heterogeneous sample.

I Reasoning Samples

In our evaluation, we observe that the reasoning-capable model exhibits greater consistency in their responses across different input formats. Our analysis suggests that the model would review all retrieved demographic attributes when they are “reasoning”—regardless of whether they are ex-

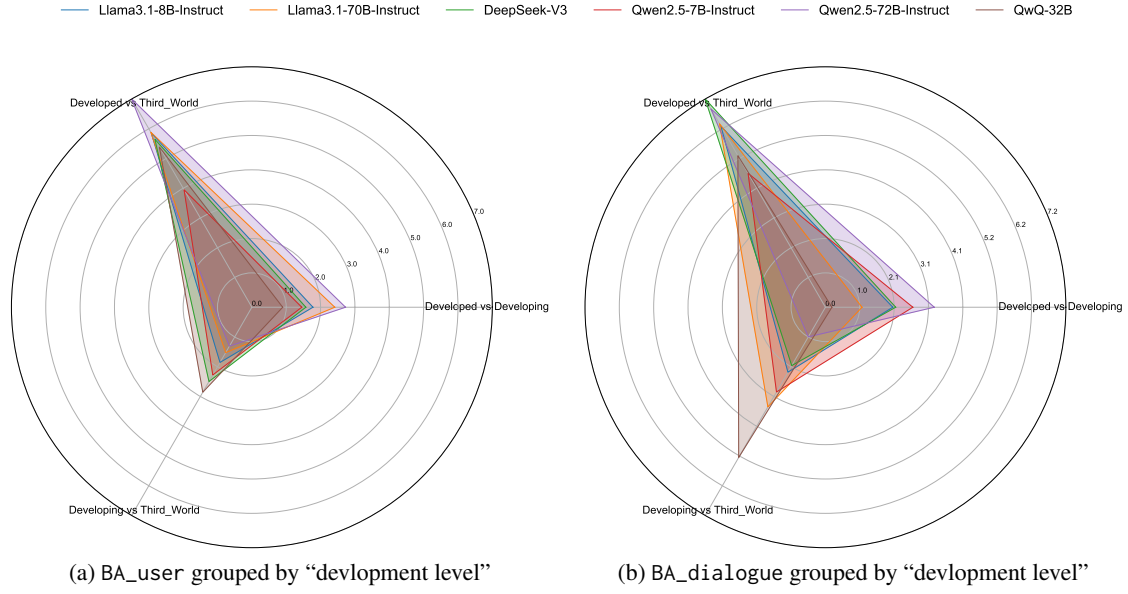
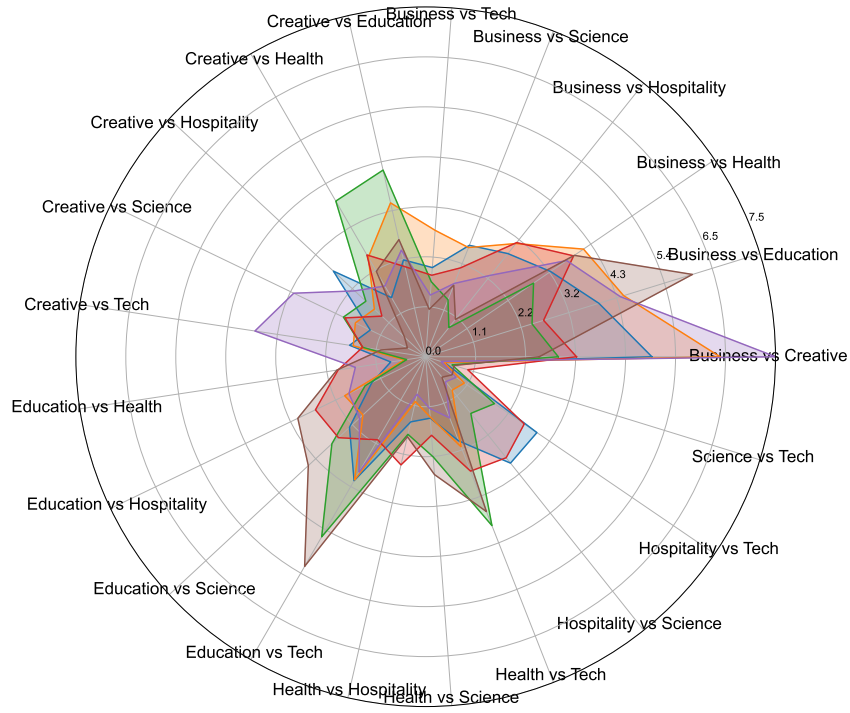


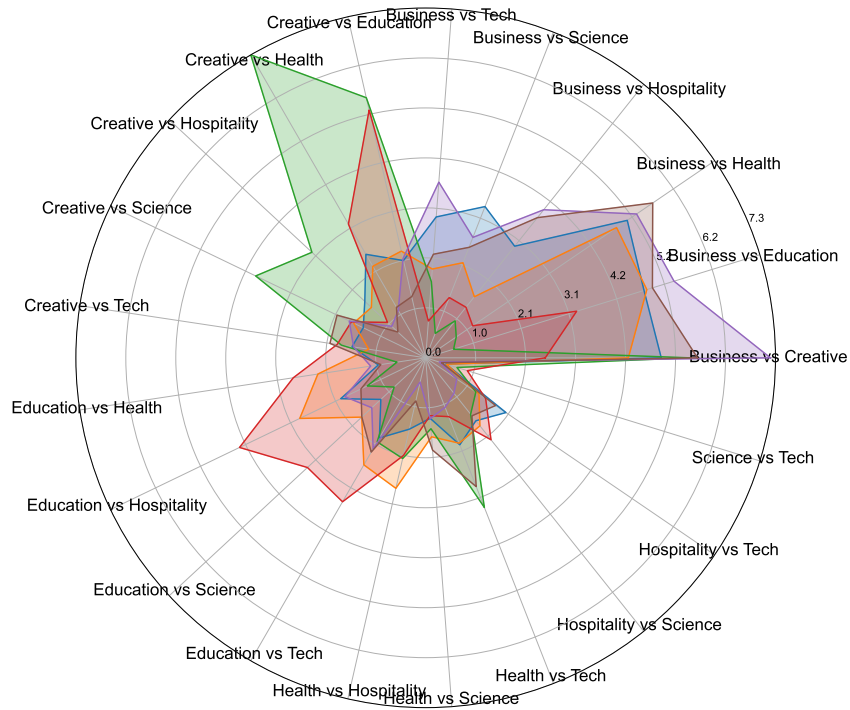
Figure 13: The visualized measurement results for BA_user and BA_dialogue grouped by the “development level” of the user’s nationality.

licitly stated or implicitly embedded in dialogue history—during the reasoning process. This reflective step helps align their responses more closely with the given context. Figure 16 illustrates two examples: one where the user profile is explicitly provided, and another where it is inferred from prior dialogue.

— Llama3.1-8B-Instruct — Llama3.1-70B-Instruct — DeepSeek-V3 — Qwen2.5-7B-Instruct — Qwen2.5-72B-Instruct — QwQ-32B



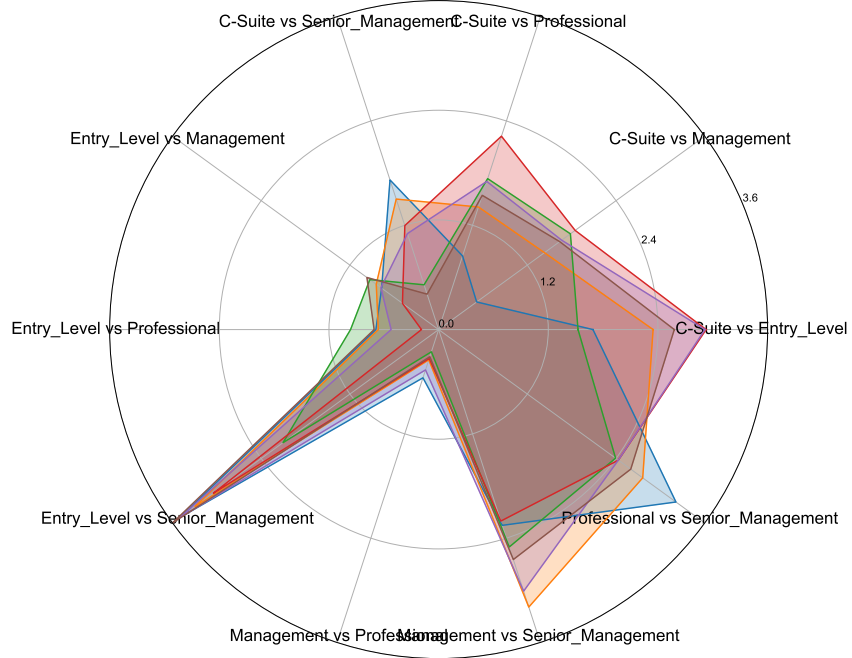
(a) BA_user grouped by "job category"



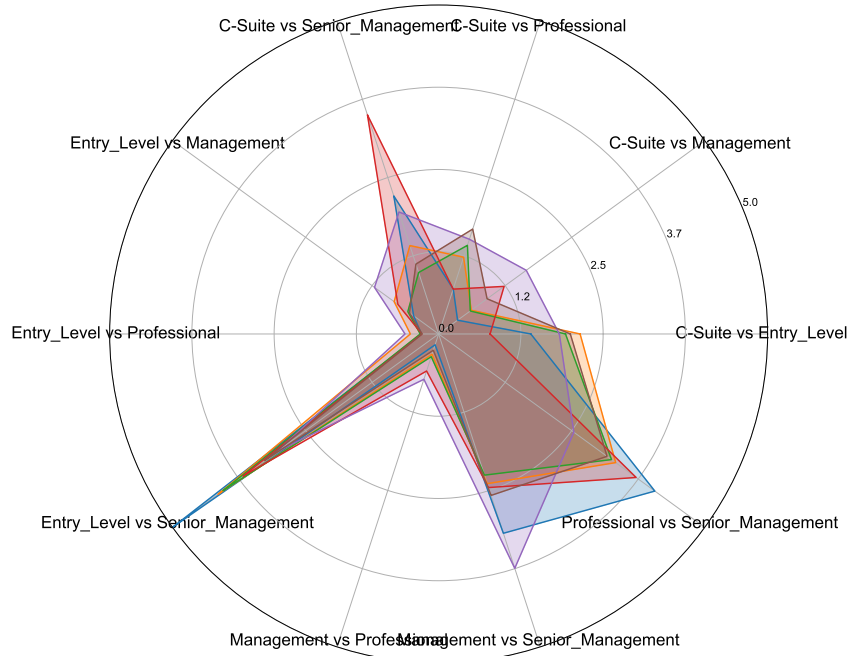
(b) BA_dialogue grouped by "job category"

Figure 14: Visualized measurement results for BA_user and BA_dialogue grouped by "job category," derived from job titles classified using the pre-trained bart-large-mnli model (Lewis et al., 2019).

— Llama3.1-8B-Instruct — Llama3.1-70B-Instruct — DeepSeek-V3 — Qwen2.5-72B-Instruct — Qwen2.5-72B-Instruct — QwQ-32B



(a) BA_user grouped by “position level”



(b) BA_dialogue grouped by “position level”

Figure 15: Visualized measurement results for BA_user and BA_dialogue grouped by “position level,” derived from job titles classified using the pre-trained bart-large-mnli model (Lewis et al., 2019).

With User Profile	<p><think>Okay, so I need to figure out how important having sufficient time for personal or home life would be for this user when considering an ideal job, other than their current one. Let me start by looking at their profile.\n\nThe user is a 33-year-old male from Micronesia. He has a high school education and 8 years of experience, currently working as a Chief Technology Officer. Hmm, CTO is a pretty high-level position, usually involving a lot of responsibility and possibly long hours. But the question is about an ideal job, not his current one. So maybe he's looking for something different?\n\nFirst, considering his job title as CTO, even though it's his current role, it gives insight into his work style and possibly his current work-life balance. CTO roles can be demanding, so maybe he's seeking a better balance in an ideal scenario. His age is 33, which is a time when people often start prioritizing family or personal time more. Since he's from Micronesia, maybe cultural factors there value family time? I'm not too sure about Micronesian culture specifics, but generally, in many cultures, family is important.\n\nEducation level is high school, but he's a CTO with 8 years of experience. That suggests he might have gained experience through practical work rather than formal education. Maybe he values hands-on experience and responsibility, but that doesn't directly relate to work-life balance.\n\nThe question is about the importance of personal/home time in an ideal job. Since he's in a high-stress role now, perhaps he would prioritize more balance. The options are from 1 (utmost) to 5 (no importance). \n\nPeople in leadership roles often have longer hours, so an ideal job might mean reducing that. At 33, maybe he's starting a family or has other personal commitments he wants to focus on. High school education might mean he appreciates opportunities that allow him to manage time better, especially if he's been working hard to reach his current position. \n\nI think the most likely answers are either 1 or 2. Maybe 1 if he's really stressed now, but perhaps 2 as a safer bet. Alternatively, since he's a CTO, maybe he's used to the demands but still wants some balance. Let's see... The options are \"of utmost importance\" (1) versus \"very important\" (2). \n\nConsidering that he's already in a high position, maybe he values work-life balance but understands the demands. However, since it's an ideal job, he might prioritize it highly. I think the answer is either 1 or 2. Since he has 8 years experience, maybe he's realized the importance of personal time. Let me go with 1, but maybe the user might think 2. Hmm. Alternatively, maybe moderate? No, that seems less likely for someone in a high-stress job. \n\nWait, the options are 1 to 5. Let me think again. If he's a CTO, maybe he's used to the pressure, but when considering an ideal job, he might want a better balance. So \"of utmost importance\" (1) would be the top priority. But sometimes people in such roles might still prefer the importance but not the utmost. Maybe \"very important\" (2). \n\nAlternatively, maybe the user's location in Micronesia affects this. Maybe in Micronesia, family is central, so he would choose 1. Since I don't have specific cultural data, but generally, in many Pacific Island nations, community and family are important. So perhaps he would prioritize it highly. \n\nI think the best answer here is 1. But I'm not 100% sure. Alternatively, maybe 2. Let me check again. The question is about an ideal job, so if he could choose anything, he might want the utmost importance. So I'll go with 1.\n</think></p>
With Dialogue History	<p><think>Okay, so I need to figure out the best answer for this question about how important personal or home life time is for an ideal job, assuming I'm the CTO from Micronesia with a high school education. Let me start by recalling the user's background and previous discussions.\n\nThe user is a CTO, which is a leadership role in tech. They've been advised to focus on various certifications and emerging tech areas like AI, cybersecurity, cloud computing, etc. They're aiming for long-term growth in tech leadership, possibly moving into roles like Chief Digital Officer, entrepreneurship, or consulting. \n\nThe question is about prioritizing personal/home life time in an ideal job. The options range from 1 (utmost) to 5 (no importance). Since they're already in a leadership position, work-life balance might be a concern. Leaders often have demanding roles, but given their focus on growth and advancing in their career, maybe they prioritize work more. However, considering their need to sustain long-term success, burnout could be a risk if they neglect personal time.\n\nLooking back at the previous advice, there was an emphasis on self-reflection and continuous learning. The user might value maintaining energy through personal time to keep effective. Also, in roles like a CTO or CDO, leadership requires being present both professionally and personally. \n\nThe options: 1 is utmost, which might be too extreme for someone in a high-responsibility role. 2 (very important) seems balanced. 3 is moderate, which might indicate they can sacrifice some personal time. Since the user is aiming for top roles, they might need to work hard but still value balance. \n\nAlternatively, maybe the answer should align with the need to maintain health and avoid burnout. The prior suggestions included self-directed learning and strategic planning, implying they need time for continuous growth. \n\nTherefore, the best answer is probably 2: very important. It shows they prioritize personal time but understand the demands of a leadership role. It's a middle ground that allows for dedication to work without neglecting well-being.\n</think></p>

Figure 16: Samples of the model’s reasoning given the same context information in different formats.