Adaptive Interviewing for Persona Simulation in LLMs: Evidence-Grounded Reasoning Improves Decision Alignment

Anonymous ACL submission

Abstract

Accurately simulating the decisions of a specific individual remains challenging for large language models (LLMs), partly because persona information is often provided as static descriptions that miss the values, experiences, and contextual cues needed for individual-level decision simulation. We propose an adaptive interview framework that gathers persona-relevant information through a structured three-stage dialogue: core questions, dynamic follow-ups, and a synthesized personality summary. Using the resulting interview transcripts, we evaluate whether LLMs can simulate participants' decisions in moral dilemma scenarios. We compare three conversational contexts—Core-10 responses, the full interview dialogue, and a summarized persona representation. We find that adaptive interviewing functions less as a uniform accuracy booster and more as a *selective grounding mechanism*: follow-up-derived evidence is incorporated in around 40% of full-interview traces, and these follow-up-grounded predictions are more accurate than core-only grounded ones (45.5% vs. 39.3%). These findings highlight that richer persona context alone is insufficient: improvements arise only when models actually ground their decisions in user-specific evidence.

1 Introduction

Large language models (LLMs) have shown increasing potential in simulating aspects of human behavior, including opinions, preferences, and decision-making patterns. Recent work suggests that LLMs can reproduce population-level opinion distributions when conditioned on demographic prompts (Argyle et al., 2023). However, accurately simulating the decisions of a specific individual remains challenging, as human decisions often depend on personal values, experiences, and contextual factors that are not captured by coarse persona descriptions.

A common approach to persona simulation is to condition LLMs on static persona attributes such as demographic profiles or short persona descriptions. While this strategy can provide useful context, prior work shows that such conditioning often fails to reliably reproduce the opinions of specific groups (Santurkar et al., 2023). In realistic settings, persona information is rarely fully specified in advance and must instead be elicited through interaction.

In this work, we study *adaptive interviewing* as a mechanism for eliciting persona information. Instead of relying solely on a fixed set of questions, our framework elicits persona information through a structured three-stage dialogue: ten open-ended questions spanning psychologically grounded domains, dynamically generated follow-up questions conditioned on prior responses, and a personality summary distilled from the full conversation.

We evaluate this approach in a decision simulation setting based on moral dilemma scenarios (Haidt, 2001; Hendrycks et al., 2021). Beyond prediction accuracy, we analyze model reasoning to examine whether decisions are grounded in evidence derived from conversational context. Our results show that while adaptive follow-up questions provide additional persona-relevant information, improvements in prediction accuracy are conditional. Instead, performance gains occur primarily when models explicitly incorporate follow-up-derived evidence into their reasoning. This finding suggests that the effectiveness of adaptive interviewing depends not only on collecting richer persona information, but also on whether models actually use that information to generate decisions.

Contributions. This paper makes three main contributions:

- We introduce an adaptive interviewing framework for persona elicitation, in which follow-up questions are dynamically generated based

on a user's previous responses to gather richer decision-relevant information.

- We propose an evaluation setup for individual decision simulation using moral dilemma scenarios and compare model performance under different conversational contexts.
- We conduct a reasoning-trace analysis on 340 categorical decision items, showing that adaptive follow-up acts as a *selective grounding mechanism*: follow-up-derived evidence is incorporated in 40% of full-interview traces, especially in constraint-based reasoning, and these follow-up-grounded predictions show higher accuracy than core-only grounded predictions (45.5% vs. 39.3%).

2 Related Works

LLMs as simulators of human opinions and personas. Recent work suggests that large language models can approximate aspects of human opinions and personas when conditioned on demographic or role-based prompts. Prior studies show that LLMs can reproduce population-level opinion patterns under demographic conditioning, indicating a baseline capacity for opinion simulation (Argyle et al., 2023). However, subsequent work finds that such simulations often remain misaligned with the opinions of specific demographic groups even when models are explicitly steered toward those identities (Santurkar et al., 2023). These findings suggest that coarse demographic labels may be insufficient for capturing the richer information required for individual-level simulation. Other work demonstrates that richer persona specifications can produce coherent social behaviors in generative agents (Park et al., 2023), while studies of LLM personality measurement suggest that model behavior is sensitive to how persona information is specified and prompted (Safdari et al., 2023). Together, these results motivate approaches that move beyond static persona descriptors toward richer, interaction-driven persona elicitation.

Persona information in dialogue systems. Dialogue research similarly highlights the importance of persona information while typically assuming that such information is already available. Persona-conditioned dialogue models (Li et al., 2016) and the PersonaChat dataset show that providing explicit persona facts can improve conversational consistency and specificity (Zhang et al., 2018). However, these approaches generally rely on predefined persona descriptions rather than dynamically eliciting them during interaction. Conversational information-seeking research distinguishes between slot-filling dialogue and exploratory conversational search (Zamani et al., 2022). Our adaptive follow-up design aligns with the latter perspective: instead of filling predefined slots, follow-up questions aim to uncover latent values, experiences, and decision-relevant cues that may not appear in initial responses.

Reasoning transparency and evidence grounding in LLMs. Recent work on reasoning transparency makes it possible to analyze how LLMs arrive at decisions. Chain-of-thought prompting improves performance on complex reasoning tasks by eliciting intermediate reasoning steps in natural language (Wei et al., 2022). Retrieval-augmented generation further shows that grounding model outputs in relevant evidence improves reliability and factual accuracy (Lewis et al., 2020). However, most LLM evaluations focus on aggregate task performance rather than examining whether model predictions are grounded in evidence derived from conversational context. In this work, we extend evaluation beyond prediction accuracy by analyzing reasoning traces and tracking the sources of evidence used in model decisions.

Value-laden decision tasks. We study these questions in moral and value-laden decision settings because such judgments depend strongly on how individuals prioritize competing values. Research in moral psychology suggests that explicit justifications often reflect post-hoc rationalization rather than direct access to internal reasoning processes (Haidt, 2001). At the same time, benchmarks evaluating ethical reasoning show that such scenarios remain challenging for language models (Hendrycks et al., 2021). We therefore use moral dilemmas not primarily to test moral knowledge, but as a demanding setting for evaluating whether richer persona elicitation leads to more user-grounded decision simulation.

3 Method

Our goal is to study whether adaptive interviewing improves the ability of large language models (LLMs) to simulate individual decision-making. Rather than relying on static persona descriptions, our framework elicits persona-relevant information

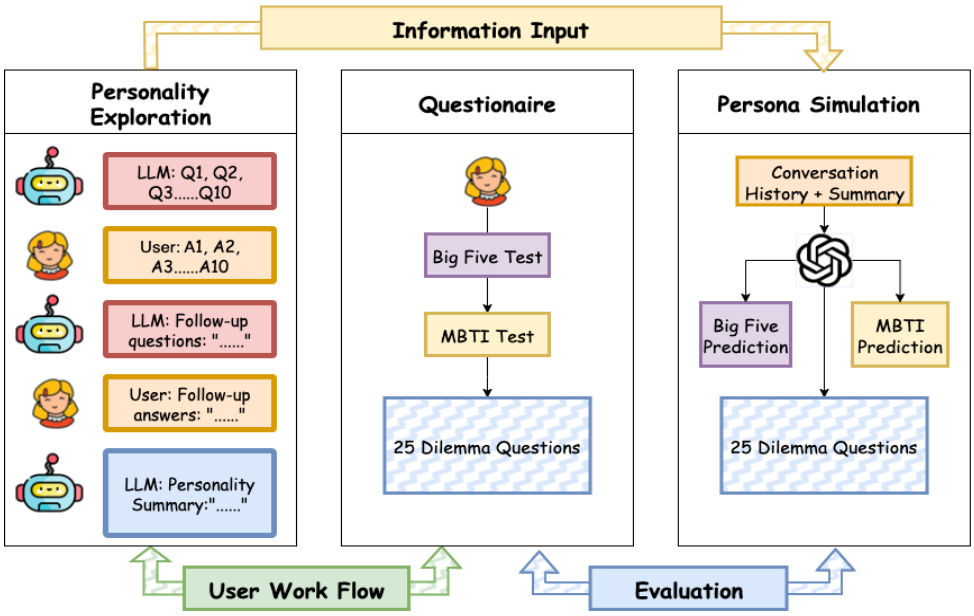


Figure 1: Overview of our method pipeline. Stage 1 elicits open-ended responses and follow-ups across 10 persona domains. In Stage 2, users complete standardized assessments (Big Five, MBTI, and 25 dilemma questions). Stage 3 evaluates the LLM's ability to simulate user traits and decisions based on prior conversation or summary by comparing to ground truth collected in Stage 2.

through multi-turn dialogue and evaluates whether this information helps predict a user's decisions in downstream tasks.

3.1 Interview Framework

We propose a three-stage adaptive persona interview for eliciting psychologically rich self-descriptions that can later be used to test an LLM's ability to infer and simulate personality. Rather than fixed questionnaires, the framework operates a semi-structured conversation: the model first poses ten open-ended questions spanning complementary psychological domains as in table 2, then, after the participant answers, it generates a small set of adaptive follow-up questions to clarify or deepen unclear aspects. Finally, the collected dialogue is used to evaluate whether an LLM can simulate the user's decisions in unseen scenarios.

All prompts and outputs are in natural language, allowing direct integration with generative models such as GPT-5 (OpenAI, 2025) and DeepSeek R1 (Guo et al., 2025). Figure 1 illustrates the overall pipeline of our method, which we detail in the following subsections.

3.2 Prompt Design and Dialogue Structure

We implement the interview using a single meta-prompt that orchestrates three stages of interaction: core question generation, adaptive follow-up, and reflective synthesis (summary). The temperature parameter (0.8–1.0) encourages variation in wording across participants while preserving semantic alignment. The full prompt template is provided in Appendix A.3.

Ten Persona Domains. The initial ten questions are constrained to cover ten empirically grounded domains drawn from personality and self-narrative psychology (Table 2). Each domain targets a different aspect of personal experience, ranging from behavioral habits and coping styles to moral conflict and life meaning. Questions are phrased to invite concrete examples ("Tell me about . . . ") rather than abstract self-assessment.

Adaptive Follow-Up Generation. After collecting the ten primary responses, the model generates five to six adaptive follow-up questions based on the earlier answers. These questions target information gaps, ambiguities, or emotionally salient segments and often reference prior statements directly (e.g., "You mentioned X . . . could you tell me more about Y?"). The follow-ups expand the dialogue and provide additional persona-relevant information for downstream reasoning tasks.

3.3 Input Settings for Downstream Evaluation

In the current experiments we do not distill each conversation into a fixed "persona card." Instead, different portions of the interview transcript are

used as contextual input for the downstream reasoning task. We compare three conversational contexts that vary in how much interview information is available to the model.

Core-10 (Baseline). The model receives only the participant's responses to the ten core interview questions. These responses provide broad coverage of the predefined persona domains but do not include adaptive follow-up dialogue.

Full Interview. The model receives the complete interview transcript, including both the Core-10 responses and the adaptive follow-up exchanges generated during the interview process.

Personality Summary. The model receives a concise personality summary automatically distilled from the full interview dialogue.

For GPT-5 we evaluate all three settings in order to study how additional conversational detail influences decision simulation accuracy. DeepSeek-R1 is evaluated under the Full Interview and Conversation Summary settings.

3.4 Evaluation Tasks

We evaluate whether the interview dialogue provides sufficient information for LLMs to simulate a participant's decisions in new contexts. Our primary task focuses on behavioral simulation, while personality inference serves as an auxiliary diagnostic.

Behavioral simulation (primary task). Given the interview context, the model answers a set of dilemma questions drawn from moral and social-decision paradigms (e.g., CNI-style trolley (Gawronski et al., 2017), altruism vs. loyalty (Haidt and Graham, 2007)). Performance is measured by how closely simulated answers match the participant's real choices.

Personality inference (auxiliary task). As an additional diagnostic, we test whether the interview dialogue also supports canonical personality prediction. Given the same interview transcript, the model predicts the participant's MBTI type (Myers et al., 1998) and continuous Big Five scores (John and Srivastava, 1999) (1–40 scale). Predictions are evaluated against participant self-reports, noting that some participants reported two possible MBTI types to reflect uncertainty along certain trait dimensions.

We report multiple evaluation metrics, including exact match accuracy, relaxed matching (e.g., hit@2, off-by-1), each tailored to different task formats. Together, these evaluations assess both decision simulation and personality inference, allowing us to examine whether conversational persona elicitation supports downstream behavioral generalization.

Condition	Accuracy	95% CI
Core-10	0.379	[0.337, 0.420]
Full Interview	0.365	[0.333, 0.402]
Personality Summary	**0.393**	[0.350, 0.433]

Table 1: Overall 25-question accuracy with 95% bootstrap confidence intervals across the three input conditions. Confidence intervals overlap substantially, indicating that aggregate accuracy differences are not statistically distinct; the mechanism-level analysis in Section 4.3 provides stronger evidence about when follow-up evidence helps.

4 Results

Experimental Setup. We conducted the study with 20 participants (balanced by gender, aged 20–30), recruited under anonymous conditions. All participants were informed of data privacy protocols and encouraged to provide genuine, self-reflective responses. All participants completed the full three-stage interaction, including the adaptive interview and follow-up questionnaires. Ethical considerations and data handling procedures are detailed in the Ethics Statement. We evaluated the collected data using the tasks outlined in Section 3.4.

4.1 Overall Performance

We summarize overall performance for dilemma prediction, MBTI recovery, and Big Five matching in Tables C1–C4; metric definitions appear in Appendix B.1. Perhaps counterintuitively, aggregate accuracy does not monotonically increase with conversational richness—we unpack this pattern in Section 4.2.

Table 1 reports overall accuracy across the three interview conditions, together with 95% bootstrap confidence intervals (bootstrap methodology detailed in Appendix C.2). The intervals overlap substantially across all three conditions, indicating that the aggregate accuracy differences should be interpreted as descriptive rather than statistically distinct. We therefore turn to a mechanism-level reasoning analysis (Section 4.3) to characterize when richer interview context actually contributes to prediction. Among the three conditions, the Personality Summary condition achieves the highest aggregate accuracy (0.393), followed by the Core-

10 interview (0.379) and the Full interview condition (0.365), suggesting that condensed summaries of conversational content may provide a compact representation of persona-relevant signals that can reduce noise in some prediction settings.

Under strict exact-match evaluation, the full interview condition does not uniformly outperform the core-only baseline. However, this aggregate metric masks substantial heterogeneity across task types and evaluation criteria. In particular, it obscures improvements in ordinal calibration under richer context.

For Likert-style questions, we additionally report an off-by-one tolerance metric to capture near-miss alignment on the ordinal scale (Table C3). As we show in Section 4.2, allowing such tolerance reveals a different pattern across interview conditions, suggesting that richer conversational context may improve ordinal calibration even when exact-match accuracy declines.

Exact-match accuracy alone does not fully capture persona simulation quality. Taken together, these aggregate results suggest that strict exact-match accuracy alone is insufficient for evaluating persona simulation quality, motivating a more fine-grained analysis across question types.

4.2 Performance by Question Type

To better understand these differences, Figure 2a breaks down performance by question type.

Across all conditions, Ranking questions achieve the highest accuracy (0.675–0.730), substantially above chance. Performance is slightly higher under the Full interview condition, suggesting that relative preference ordering may benefit from richer conversational context, potentially because such tasks depend on trade-offs and value priorities.

Choice questions show similar performance across conditions (0.391–0.403), indicating limited sensitivity to representation format compared to other task types.

In contrast, Likert-style questions exhibit larger variation across conditions. Because Likert responses represent ordinal judgments, we also report an off-by-one tolerance metric that counts predictions within one scale point of the ground truth as near-miss alignment (Table C3). Under strict exact-match evaluation, the Personality Summary condition achieves the highest accuracy (0.329), followed by Core-10 (0.307), while the Full interview condition performs lowest (0.236). However, when allowing a one-point tolerance, the pattern reverses: the Full interview condition achieves substantially stronger ordinal calibration (74.3% vs. 57.9%). This suggests that richer conversational context helps the model place participants near the correct region of the response scale, even when the exact discrete point is missed.

Overall, these results indicate that **different persona simulation tasks benefit from different levels of contextual compression**. Tasks requiring discrete categorical judgments (e.g., multiple-choice) perform best with condensed persona summaries, while tasks involving ordinal calibration or preference trade-offs benefit more from richer conversational context.

The aggregate accuracy patterns above raise a key question: when does richer conversational context actually help? In Section 4.3, we analyze reasoning traces directly to show that the benefit is selective—concentrated in predictions that explicitly ground on follow-up-derived evidence.

4.3 Reasoning Evidence and the Selective Value of Adaptive Follow-up

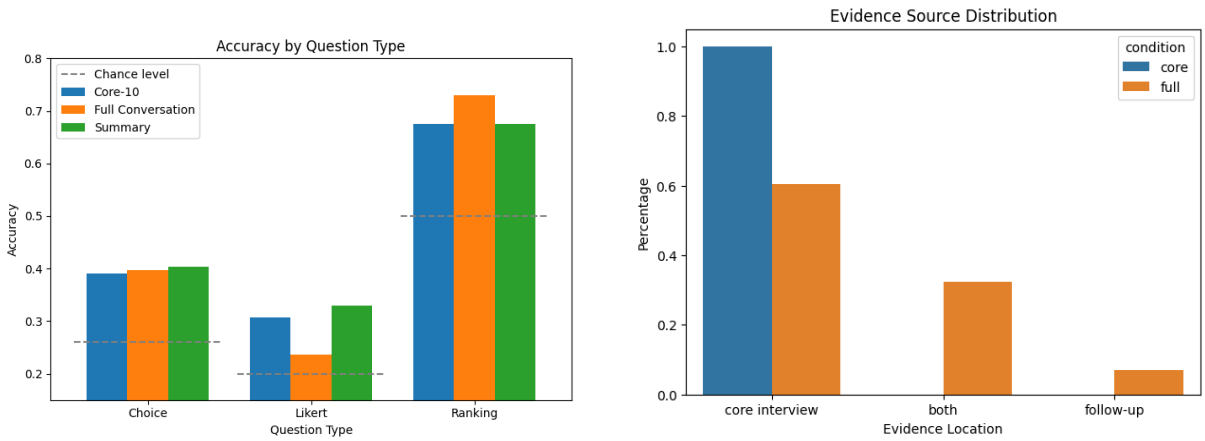
4.3.1 Overview

To better understand why richer interview context does not uniformly improve prediction accuracy, we analyze model-generated reasoning traces under the baseline (core-10 only) and full-interview conditions. For each prediction, the model outputs (1) a short natural-language explanation, (2) an evidence excerpt supporting the prediction, (3) the location of the supporting evidence (core interview vs. follow-up vs. both), and (4) a pre-annotated reasoning category.

Our analysis focuses on two complementary questions: which conversational evidence sources are used in reasoning, and whether predictions grounded in follow-up-derived evidence are more accurate. Rather than establishing causal relationships, this analysis aims to identify systematic associations between interview structure, evidence grounding, and prediction quality.

Reasoning Taxonomy. We use a coarse reasoning taxonomy as defined in table D1 to label the dominant form of explanation in each trace. Because value-based reasoning dominates in both settings, we treat reasoning category as a supporting descriptive signal rather than the main outcome of interest.

Annotation Reliability. Annotation quality was



(a) Accuracy by question type across interview conditions.

(b) Distribution of evidence sources used in model reasoning.

Figure 2: Analysis of model performance and reasoning evidence. **(a) Accuracy by question type across interview conditions.** Bars show average exact-match accuracy for Choice, Likert, and Ranking questions under the Core-10 interview, Full interview, and Personality Summary conditions. Dashed horizontal lines indicate approximate chance levels for each question type. **(b) Distribution of evidence sources used in model reasoning.** While the Core-10 condition relies exclusively on core interview information, the Full interview condition incorporates follow-up-derived evidence either alone or jointly with core interview responses.

validated through human verification, yielding 95% inter-rater agreement among annotators and 87.9% agreement with LLM pre-labels; details of the annotation procedure are provided in Appendix D.1.

4.3.2 Evidence Usage and Prediction Accuracy

Reasoning-Type Distribution Across conditions, coarse reasoning-category distributions remain broadly similar, with value-based reasoning dominating in both settings, accounting for 88.5% of Core-10 traces and 91.8% of Full-interview traces. Constraint-based and narrative reasoning remain comparatively rare. A small number of generic traces appear despite the grounding prompt, suggesting that the interview does not always contain decision-relevant evidence for every item.

Evidence Sources and Accuracy In the Full interview condition, 40% of model predictions incorporate follow-up-derived evidence, either exclusively or jointly with core interview responses. The majority of reasoning traces still rely on core interview content, indicating that follow-up evidence supplements rather than replaces the initial interview context. We also observe that Follow-up evidence is incorporated more often in constraint-based reasoning than in value-based reasoning (57.9% vs. 39.4%, see Appendix D2). This suggests that adaptive follow-up is especially useful when the model needs to reason about practical constraints, trade-offs, or contextual limitations rather than abstract values alone.

Interestingly, in the categorical-choice reasoning audit, predictions incorporating follow-up-derived evidence are more accurate than predictions grounded only in the core interview context, but the difference is modest (45.5% vs. 39.3%). Confidence intervals for these estimates are reported in Appendix D.4. This suggests that follow-up evidence provides a selective advantage when the model actually grounds its reasoning in them, rather than a large uniform gain.

Paired Prediction Analysis To further probe this effect, we compare these follow-up-grounded Full-condition predictions with their matched Core-10 counterparts for the same participant–question pairs. Among the 134 cases in which Full-condition reasoning incorporates follow-up-derived evidence, the model changes its prediction in 50.7% of cases. Among these changed predictions, improvements (15 cases) substantially outnumber degradations (6 cases), while the remaining changed cases stay incorrect under both settings. This paired comparison suggests that when follow-up information alters the model's prediction, it is more likely to correct an earlier mistake than to introduce a new one. Detailed transition counts are reported in Appendix D3.

The benefit of adaptive follow-up is selective and evidence-dependent. Richer interview context

does not uniformly improve all predictions. Instead, its value becomes visible when the model meaningfully uses the additional cues elicited through adaptive follow-up. Taken together, the evidence-source analysis and the paired comparison above help explain why full-interview input does not always improve aggregate exact-match accuracy, even though follow-up-grounded cases are more accurate and prediction changes are more often beneficial than harmful.

4.4 Qualitative Error Analysis: Default Interpretive Biases

To better understand why richer conversational context does not always yield correct predictions, we examine representative failure cases and identify several recurring interpretive biases in model reasoning.

Normative Rule Bias LLMs frequently default to rule-following interpretations even when users express more pragmatic or contextual decision strategies. For example, in Q21 (a legal scenario involving insufficient evidence), models consistently predicted strict adherence to legal procedure—releasing the suspect and continuing investigation—while some users favored more cautious or situational interpretations of responsibility. Similarly, in Q25, models often prioritized informational completeness and procedural fairness even when the user's reference choice emphasized practical trade-offs.

Prosocial Harmony Bias LLMs also show a tendency to favor cooperative or harmony-preserving actions. In Q2, which asks how one responds to a friend repeatedly cancelling plans, models frequently selected communication- or problem-solving-oriented responses, assuming that the user would attempt dialogue or relationship repair. However, some participants instead preferred disengagement or boundary-setting. A similar pattern appears in Q22, where models often predicted inclusive activity choices that accommodate everyone, reflecting a preference for socially harmonious outcomes.

Over-responsibility Bias Finally, models sometimes favor compromise-oriented solutions that attempt to satisfy multiple competing obligations simultaneously. In Q24, LLM predictions often selected "two-win" strategies even when users ultimately chose a single pragmatic option due to real-world constraints. This pattern suggests that models may overestimate individuals' willingness to pursue ideal conflict resolution rather than realistic trade-offs.

Taken together, these patterns suggest that prediction errors arise not only from missing evidence but also from systematic interpretive biases in how models resolve ambiguous social scenarios. These tendencies align with prior observations that aligned language models often default toward norm-consistent, cooperative, and socially acceptable responses (Ouyang et al., 2022; Sharma et al., 2024; Scherrer et al., 2023; Samway et al., 2025; Park et al., 2025).

More broadly, these findings highlight the challenge of persona modeling in contexts where individuals may deviate from normative or majority-aligned decision patterns. Combined with the reasoning evidence analysis in Section 4.3, this suggests that improving persona simulation requires not only richer evidence elicitation but also mechanisms that encourage models to ground predictions in user-specific signals rather than default normative heuristics.

4.5 MBTI and Big-Five Predictions

As an auxiliary personality-recovery evaluation, we also compare interview conditions on MBTI and Big Five prediction. MBTI prediction results show broader plausible type recovery under the Full interview condition than under Core-10 (0.40 vs. 0.20 for $hit@2$), while the Personality Summary condition performs best overall (0.50). This pattern is consistent with our earlier findings: high-level typological judgments benefit from compressed persona representations, whereas fuller conversational context can still improve approximate personality recovery. Detailed MBTI and Big Five results are reported in Appendix C.

5 Discussion

5.1 "More Context = Better Simulation?"

Although follow-up-grounded predictions are more accurate (45.5% vs. 39.3%), aggregate accuracy under the Full interview condition does not consistently exceed the Core-10 baseline. We interpret this as a selective grounding dynamic: follow-up questions increase the availability of persona-relevant cues, but prediction accuracy improves only when those cues are actually incorporated into reasoning. In this sense, adaptive follow-up acts

less as a direct accuracy booster and more as an **information-availability mechanism**. By probing clarifications, motivations, and higher-level reflections, follow-up questions surface additional persona signals that may remain latent in static questionnaires. However, the downstream benefit of these signals depends on whether the model meaningfully grounds its predictions in them.

5.2 Task-Adaptive Persona Representations

Our task-type analysis in Section 4.2 suggests a practical design principle: the optimal persona representation depends on the downstream task. For categorical decisions (multiple-choice dilemmas, MBTI inference), condensed personality summaries perform better by filtering redundant detail. For ordinal judgments (Likert-scale items) and preference ranking tasks, richer conversational context supports finer calibration. These results indicate that persona systems should avoid uniform compression; instead, task-adaptive representations—summarized for discrete tasks and full-history for ordinal tasks—may better support diverse downstream applications.

5.3 Implications and Future Directions

These findings highlight two broader implications for persona modeling. First, richer elicitation protocols alone are insufficient; improvements depend on whether downstream inference mechanisms reliably **ground predictions in elicited evidence**. Second, evaluation of persona simulation should consider both **predictive alignment** and **reasoning grounding**, rather than relying solely on strict exact-match accuracy.

Future work may explore follow-up question designs that more consistently surface decision-relevant persona cues, develop richer reasoning annotations to capture conversational evidence, and investigate architectures that better incorporate multi-turn persona information into downstream inference. More systematic analysis of user-level moderators may also clarify when adaptive elicitation yields the greatest benefits.

6 Conclusion

We introduce an adaptive interviewing framework that elicits persona-relevant information through a structured three-stage dialogue—core questions, dynamic follow-up, and summary—and evaluate its impact on LLM-based decision simulation.

Our results yield two main findings. First, different persona simulation tasks benefit from different levels of contextual compression: condensed summaries perform best for categorical predictions (multiple-choice dilemmas, MBTI inference), while richer conversational context better supports ordinal calibration in Likert-scale judgments and ranking tasks. Second, the benefit of adaptive follow-up is selective rather than uniform: follow-up-derived evidence is incorporated in roughly 40% of Full-interview traces, and these follow-up-grounded predictions show modestly higher accuracy than core-only grounded predictions on categorical choice items (45.5% v.s. 39.3%). Together, these findings suggest that effective persona simulation requires not only richer elicitation, but also reliable grounding of model reasoning in user-specific evidence.

Limitations

Our study is subject to several limitations.
First, our participant sample (N=20) is modest in size and relatively homogeneous (primarily aged 20–30), which limits statistical power and may constrain generalizability across broader populations and moral reasoning styles. The reported accuracy differences across interview conditions should be interpreted as descriptive given overlapping confidence intervals; we therefore frame our findings as evidence for a *selective grounding mechanism* rather than a uniform accuracy effect.
Second, the fixed ordering of interview followed by evaluation may introduce priming effects, where participants' reflections during the interview influence their subsequent responses to dilemma questions.
Third, our pipeline involves LLMs at multiple stages, including question generation, follow-up prompting, summary construction, and downstream prediction. This raises the possibility that LLM-generated summaries encode persona information in a format that is particularly amenable to LLM-based inference, potentially contributing to the strong performance of the Summary condition in categorical tasks. A preliminary evaluation on DeepSeek-R1 under the Full Interview condition yields comparable overall accuracy (0.385 vs. 0.365 for GPT-5), suggesting that the patterns observed here are not specific to a single model family; however, a systematic cross-model analysis—including reasoning-trace audits on additional models—is left to future work.
Fourth, while our dilemma set spans multiple thematic domains (Table C5)—moral reasoning, social cooperation, emotion regulation, decision-making, value trade-offs, and personality-relevant scenarios—it remains centered on value-laden judgments. Decision contexts that emphasize probabilistic reasoning under uncertainty (e.g., risk and reward trade-offs) or low-stakes everyday preferences (e.g., consumption choices) are not explicitly represented, and extending the framework to such settings is a natural direction for future work.

Ethics Statement

Ethical Compliance. This study did not undergo formal IRB review. Our institution does not require IRB review for this research, as it involves only voluntary adult participants, employs no deception, collects fully anonymized data, and does not involve sensitive or protected personal attributes. The study design is consistent with the criteria for exemption under standard human subjects research guidelines (45 CFR 46.104, Category 2).

Participant Recruitment and Consent. Participants were recruited via online academic networks. All participants were adults who volunteered without monetary compensation for a study lasting approximately 60 minutes. Prior to participation, all participants provided informed consent and were briefed on the study's purpose, the scope of data collection, the anonymized use of their responses solely for NLP research, and their right to withdraw at any time without consequence. No deception was employed at any stage of the study.

Data Privacy and Anonymization. All interview transcripts were stored using anonymized participant identifiers (P01–P20); no personally identifying information was retained. Participants were instructed not to include identifying details in their responses. Transcripts are stored on encrypted local storage and will not be publicly released. A representative subset was reviewed for offensive or harmful content; none was identified.

Potential Risks and Deployment Limitations. This research involves several potential risks. First, adaptive interviewing systems that elicit detailed personal narratives could be misused for psychological profiling or surveillance without participants' knowledge. Second, LLM-based persona simulation, if deployed prematurely, could lead to incorrect behavioral inferences in consequential settings (e.g., hiring, clinical assessment). Third, the normative interpretive biases identified in Section 4.4 suggest that deployed systems might systematically misrepresent individuals who deviate from majority-aligned behavioral patterns. We emphasize that this work is a research prototype and should not be deployed in real-world decision-support contexts without further validation.

References

Gordon W. Allport. 1937. *Personality: A Psychological Interpretation*. Holt.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Chris Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Jonathan Baron. 1994. *Thinking and deciding*. Cambridge University Press.

Kim Bartholomew and Leonard M Horowitz. 1991. Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology*, 61(2):226–244.

John Bowlby. 1988. *A secure base: Parent-child attachment and healthy human development*. Basic Books.

Jennifer K. Connor-Smith and Celeste Flachsbart. 2007. Relations between personality and coping: A meta-analysis. *Journal of Personality and Social Psychology*, 93(6):1080–1107.

Paul T. Costa and Robert R. McCrae. 1992. *NEO PI-R Professional Manual*. Psychological Assessment Resources.

Bertram Gawronski, Joshua Armstrong, Paul Conway, Rebecca Friesdorf, and Mandy Hütter. 2017. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, 113(3):343–376.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34.

Jeffrey A Gray. 1987. *The psychology of fear and stress*, 2nd edition. Cambridge University Press.

James J Gross. 2002. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in large language models via reinforcement learning. *arXiv preprint arXiv:2501.12948*. Available at https://arxiv.org/abs/2501.12948.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. ArXiv:2008.02275.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press, New York, NY.

Steinar Kvale and Svend Brinkmann. 2009. *InterViews: Learning the Craft of Qualitative Research Interviewing*, 2nd edition. Sage.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 994–1003, Berlin, Germany.

Dan P McAdams. 2006. The redemptive self: Generativity and the stories americans live by. *Research in Human Development*, 3(2-3):81–100.

Dan P. McAdams and Kate C. McLean. 2013. Narrative identity. *Current Directions in Psychological Science*, 22(3):233–238.

Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Matthew B. Miles, A. Michael Huberman, and Johnny Saldãna. 2014. *Qualitative Data Analysis: A Methods Sourcebook*, 3rd edition. Sage.

Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. 1998. *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, Palo Alto, CA.

OpenAI. 2025. GPT-5: Our most advanced model for coding and reasoning. OpenAI blog post. https://openai.com/index/introducing-gpt-5/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022.

Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 27730–27744.

Bumjin Park, Jinsil Lee, and Jaesik Choi. 2025. Deontological keyword bias: The impact of modal expressions on normative judgments of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7277–7296, Vienna, Austria. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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 2023)*.

James R Rest. 1986. *Moral development: Advances in research and theory*. Praeger.

Mustafa Safdari, Gregory Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. 2023. Personality traits in large language models. ArXiv:2307.00184.

Keenan Samway, Max Kleiman-Weiner, David Guzman Piedrahita, Rada Mihalcea, Bernhard Schölkopf, and Zhijing Jin. 2025. Are language models consequentialist or deontological moral reasoners? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30699–30726, Suzhou, China. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, Honolulu, Hawaii. ArXiv:2303.17548.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*. ArXiv:2307.09491.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*, 25:1–65.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, and 1 others. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Philip E. Tetlock. 1986. A value pluralism model of ideological reasoning. *Journal of Personality and Social Psychology*, 50(4):819–827.

Wim J. van der Linden and Peter J. Pashley. 2010. Item response theory. In Wim J. van der Linden and Cees A. W. Glas, editors, *Elements of Adaptive Testing*, pages 1–18. Springer.

Simine Vazire. 2010. Who knows what about a person? the self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98(2):281–300.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 24824–24837.

Hamed Zamani, Johanne R. Trippas, Jeffrey Dalton, and Filip Radlinski. 2022. Conversational information seeking. *Foundations and Trends in Information Retrieval*, 17(3–4):244–456.

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 (ACL 2018)*, pages 2204–2213, Melbourne, Australia.

A Interview Materials

A.1 Psychological and Methodological Foundations

Personality psychology. The Five-Factor Model provides a widely used framework for describing population-level personality variation (McCrae and John, 1992; Costa and McCrae, 1992). Trait taxonomies in personality psychology are historically grounded in the lexical hypothesis, which suggests that important individual differences are reflected in natural language descriptions (Goldberg, 1993). However, personality psychologists have long distinguished between nomothetic approaches that model population-level regularities and idiographic approaches that focus on individual-specific characteristics (Allport, 1937). Our adaptive follow-up design is conceptually aligned with this idiographic perspective, aiming to elicit individual narratives and value structures rather than relying solely on predefined personality categories.

Adaptive testing. Our questioning strategy also draws inspiration from adaptive testing in psychometrics. Computerized adaptive testing dynamically selects questions based on previously observed responses in order to reduce uncertainty about a respondent's latent traits (van der Linden and Pashley, 2010). Similarly, our adaptive follow-up questions are generated in response to earlier answers to gather information that is most informative for predicting subsequent decisions.

Qualitative research foundations. Finally, our annotation scheme is informed by traditions in qualitative research. Interview methodology distinguishes between "mining" approaches that extract predefined information and "traveler" approaches that co-construct meaning through dialogue (Kvale and Brinkmann, 2009). Our adaptive follow-up questions follow the latter perspective, aiming to reveal latent value structures through iterative interaction. Our reasoning and evidence coding framework follows established practices in qualitative data analysis and systematic coding (Miles et al., 2014).

A.2 10 Persona Domains

A.3 Meta-Prompt Used to Elicit Participant Personality Responses

```
You are a deep persona interviewer whose purpose
↪  is to understand the participant
as completely as possible — their reasoning
↪  patterns, fears, values, blind spots,
and worldview.

Your interview proceeds in three stages:
---
Stage 1 — Generate Ten Core Questions
1. Generate exactly 10 open-ended, concrete,
↪  introspective questions that—taken
   together—let you form a 360-degree picture of
   ↪  the participant.
2. Each question must correspond to one of these
↪  ten thematic domains (use each once):
   1. Behavioral Priorities & Trade-offs
   2. Decision Logic & Value Hierarchy
   3. Fears & Deep Motivation
   4. Self-Awareness & Blind Spots
   5. Stress & Coping Mechanisms
   6. Relational Logic & Boundaries
   7. Value Conflicts & Compromise
   8. Core Identity & Self-Definition
   9. Emotional Vulnerability & Regulation
   10. Life Narrative & Sense of Meaning
3. Questions must be specific and
↪  story-eliciting, avoiding abstract or yes/no
↪  phrasing.
4. Tone: natural, empathic, intellectually
↪  curious.

Output template:
Stage 1 — Ten Questions:
1. ...
2. ...
...
10. ...
---
Stage 2 — Adaptive Follow-Up (trigger after
↪  participant answers)
After the participant provides answers to all ten
↪  questions, you will:
1. Read all ten answers carefully.
2. Identify 3-5 aspects that seem unclear,
↪  internally inconsistent, emotionally charged,
   or highly characteristic.
3. For each, craft a targeted follow-up question
↪  to deepen or clarify your understanding.
4. Reference the participant's words when
↪  possible.

Output template (after answers received):
Stage 2 — Follow-Up Questions:
F1. [follow-up + brief note of purpose]
F2. ...
F3. ...
---
Interaction style
- Do not give analysis or labels before Stage 2.
- Remain open, gentle, and precise.
- When Stage 2 follow-ups are answered, you may
↪  summarize the participant's personality
  insights per domain (stage3).
```

Table 2: Ten Persona Domains

#	Domain	Psychological Basis
1	Behavioral priorities & trade-offs	Value enactment theory (Schwartz, 1992)
2	Decision logic & value hierarchy	Moral reasoning & cognitive-style literature (Rest, 1986; Baron, 1994)
3	Fears & deep motivation	Reinforcement Sensitivity Theory (Gray, 1987)
4	Self-awareness & blind spots	Self–other knowledge asymmetry (Vazire, 2010)
5	Stress & coping mechanisms	Trait–coping meta-analysis (Connor-Smith and Flachsbart, 2007)
6	Relational logic & boundaries	Attachment and interpersonal theory (Bowlby, 1988; Bartholomew and Horowitz, 1991)
7	Value conflicts & compromise	Integrative complexity & ethical decision-making (Tetlock, 1986)
8	Core identity & self-definition	Narrative identity (McAdams, 2006)
9	Emotional vulnerability & regulation	Emotion regulation framework (Gross, 2002)
10	Life narrative & sense of meaning	Meaning-making & generativity literature (McAdams and McLean, 2013)

A.4 Participant Information Sheet Summary

Before beginning the study, all participants received the following information:

1. **Study purpose.** The study investigates the ability of large language models to simulate user personality based on conversational evidence.

2. **Tasks.** Participants completed three parts: (1) a multi-turn interview with a DeepSeek reasoning model (10–16 questions) designed to elicit authentic responses about personality, behavioral motivations, and value judgments; (2) a Big Five (BFI-44) personality assessment and self-reported MBTI result; and (3) 25 scenario-based moral dilemma questions.

3. **Data collection.** The full interview transcript (questions and responses), personality assessment results, and dilemma responses were recorded via survey form.

4. **Data handling.** All data is collected anonymously, used solely for NLP research purposes, and will not be publicly released.

5. **Voluntary participation.** Participation was entirely voluntary. Participants were free to withdraw at any time and were encouraged to contact the research team with any questions.

No deception was employed at any stage of the study.

B Evaluation Materials

Dilemma Question Set. We include the full set of 25 dilemma questions that we crafted and used in our evaluation stage at the following link:
PDF: Dilemma Question Set
The dilemma set additionally includes five consistency probe pairs ((Q8, Q9), (Q11, Q12), (Q14, Q15), (Q19, Q20), (Q21, Q22)) targeting closely related underlying constructs. The probes use disguised option orderings (e.g., the equal-split option is labeled A in Q8 but C in Q9) and reverse-coded Likert framings to reduce demand characteristics. A systematic analysis of probe-based behavioral consistency is left to future work.

Big Five Personality Test. We used the 44-item Big Five Inventory (BFI-44) provided by MindWorks Counselling. Participants accessed and completed the test via:
Big-5 Personality Test

Codes for Evaluation. We provide the evaluation codes for your convenience to reproduce the experiments:
Google Colab: Evaluation Codes

Licenses and Computational Details. GPT-5 was accessed via the OpenAI API under OpenAI's Terms of Service for research use. DeepSeek-R1 is released under the MIT License. The BFI-44 personality inventory is freely available for non-commercial academic research. All LLM inference was performed via commercial API; no model training or fine-tuning was conducted. DeepSeek-R1 has 671B parameters (Guo et al., 2025); GPT-5's parameter count has not been publicly disclosed by OpenAI. The approximate total API cost was USD 30. Experiments were conducted between November 2025 and March 2026.

Reproducibility. For interview generation (core question, follow-up, and summary stages), we use temperature 0.8–1.0 to encourage variation in wording across participants while preserving semantic alignment. For downstream prediction and reasoning-trace generation, we use temperature 0 (greedy decoding) to ensure deterministic outputs. The full meta-prompt template is versioned and provided in Appendix A.3. All API calls use default `top_p` and `max_tokens` settings. Evaluation scripts (including bootstrap CI computation and reasoning-trace aggregation) are linked in Appendix B.

B.1 Evaluation Metrics

To quantitatively assess the alignment between the model-predicted persona representations and participants' self-reports, we compute several complementary metrics for both the categorical MBTI and the binary Big-Five traits, which the 1-40 scale is binarized into two bins [1, 20] and [21, 40].

Exact Match (Top-1 Accuracy). The proportion of participants whose top-1 predicted MBTI type (or Big-Five binary vector) exactly matches the ground-truth label:

$$\text{Exact} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i].$$

Top-2 Hit Rate (Hit@2). Since the model outputs two most probable MBTI types, this metric counts a prediction as correct if either of the top-2 candidates appears in the participant's self-reported type set:

$$\text{Hit@2} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_{i,1} \in Y_i \text{ or } \hat{y}_{i,2} \in Y_i].$$

Off-by-1 Rate. For MBTI and Big-Five predictions, we define *off-by-1* as cases where the predicted and true personality representations differ by exactly one dimension (e.g., ENFP → INFP, or only the *Extraversion* bit is misclassified):

$$\text{Off-by-1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{HammingDist}(\hat{y}_i, y_i) = 1].$$

Off-by-2 Rate. Analogously, *off-by-2* indicates predictions that deviate from the true label on exactly two out of the four (MBTI) or five (Big-Five) binary dimensions.

Average Misclassification Rate. We further compute the mean proportion of incorrectly predicted dimensions across all participants:

$$\text{MisclassRate} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{HammingDist}(\hat{y}_i, y_i)}{D},$$

where $D = 4$ for MBTI and $D = 5$ for Big-Five traits.

These metrics jointly capture not only exact accuracy but also the *degree of deviation* in the model's inferred persona representations, providing a finer-grained understanding of how closely the LLM internalizes user-specific personality signals.

C Results Tables

C.1 MBTI and Big Five Prediction

MBTI No single interview condition consistently dominates across all four MBTI dimensions. The History condition shows slightly fewer errors on the I/E dimension, while the Summary representation performs better on N/S and T/F and achieves stronger exact and top-2 recovery. However, given the discrete nature of MBTI labels and the limited sample size, these results should be interpreted cautiously.

Big Five Across conditions, performance differences remain relatively small. All settings show difficulty predicting *Extraversion* and *Neuroticism*, while *Openness*, *Conscientiousness*, and *Agreeableness* show comparable performance across settings with minor variations. Because the original personality scores are continuous (0–40), we discretize them into binary low/high categories for evaluation; therefore these results should be interpreted as coarse personality recovery signals rather than precise trait prediction.

C.2 25 Dilemma Questions Evaluation

Relaxed evaluation for Likert questions. Because Likert questions represent ordinal judgments, we additionally report a relaxed off-by-one evaluation that counts predictions within one scale point of the gold label as correct. Under this metric, accuracy increases substantially across all conditions (see Table C3), indicating that many errors correspond to small ordinal deviations rather than completely incorrect persona inference.

Confidence Intervals. The 95% bootstrap confidence intervals reported in Table 1 are computed by resampling participants with replacement and recomputing the overall accuracy following the same evaluation protocol as in the main text, including per-question averaging and relative-order scoring for Q17.

For the overall accuracy experiment (n = 500 instances per condition), confidence intervals remain relatively tight, indicating stable estimates, but overlap across conditions, consistent with the descriptive interpretation in Section 4.1.

Note. All values are accuracies (0–1). Ties are bolded for all co-maxima.

Table C1: MBTI prediction metrics. Higher is better for *hit_at_2* and *top1_exact*; lower is better for *off_by_1* and *off_by_2*. Best values are in **bold**.

	hit_at_2	top1_exact	off_by_1	off_by_2
DeepSeek	**0.50**	0.30	**0.30**	0.20
GPT-5 (Core-10)	0.20	0.10	0.55	0.25
GPT-5 (Full Conv.)	0.40	0.10	0.45	0.15
GPT-5 (Summary)	**0.50**	**0.40**	0.45	**0.05**

Note. Ties are bolded for all co-best values.

Table C2: Big Five match rate per dimension (higher is better). Column-wise maxima are in **bold**.

	O	C	E	A	N
DeepSeek	0.85	0.80	**0.45**	**1.00**	0.50
GPT-5 (Core-10)	0.80	0.55	0.40	0.60	**0.65**
GPT-5 (Full Conv.)	0.85	**0.85**	0.30	0.80	0.30
GPT-5 (Summary)	**0.90**	0.80	0.30	0.65	0.20

Note. O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism.

Table C3: Likert evaluation results. Exact denotes strict match accuracy; Off-by-1 counts predictions within one category of the gold label. Combined reports the union of exact and off-by-1 matches.

Condition	Exact	Off-by-1
Core-10	0.307	0.579
Full Interview	0.236	0.743
Personality Summary	0.329	0.721

Table C4: Per-question accuracy by different GPT-5 setting (higher is better). Row-wise maxima are in **bold**.

Question	Core-10	Conv. History Ctx	Conv. Summary Ctx
Q1	0.300	**0.350**	0.250
Q2	0.500	0.550	**0.650**
Q3	**0.550**	0.350	0.300
Q4	0.300	0.350	**0.400**
Q5	0.300	0.300	**0.400**
Q6	0.600	**0.650**	0.450
Q7	**0.350**	**0.350**	0.300
Q8	0.100	**0.450**	0.400
Q9	0.300	0.350	**0.500**
Q10	**0.300**	**0.300**	**0.300**
Q11	**0.450**	0.250	0.300
Q12	0.250	0.200	**0.300**
Q13	0.250	**0.300**	0.250
Q14	0.350	0.300	**0.400**
Q15	**0.500**	0.250	0.400
Q16	0.250	0.100	**0.400**
Q17	0.675	**0.730**	0.675
Q18	0.300	0.200	**0.350**
Q19	**0.150**	0.100	0.050
Q20	**0.250**	**0.250**	**0.250**
Q21	**0.450**	0.400	0.400
Q22	0.250	**0.300**	**0.300**
Q23	0.450	**0.500**	**0.500**
Q24	**0.700**	0.650	**0.700**
Q25	**0.600**	**0.600**	**0.600**

Table C5: Categorization of questions by thematic domain.

Category	**Question IDs**
Moral reasoning	4, 10, 19, 21, 25
Social cooperation & fairness	8, 9, 13, 22
Emotion regulation	2, 5, 16
Decision making	3, 14, 15
Value	6, 7, 17, 18, 20
Personality trait	23, 24

D Reasoning Analysis

The reasoning analysis focuses on categorical-choice decision items. We exclude Likert-scale and ranking items from this analysis because their evaluation requires task-specific metrics, such as ordinal distance or relative-order accuracy, rather than the binary exact-match correctness used in the reasoning audit. This allows the reasoning analysis to compare prediction correctness consistently across items.

D.1 Annotation Procedure and Scope

To analyze reasoning patterns at scale, we adopt an LLM-assisted pre-annotation and human verification protocol over the full categorical-choice subset of the evaluation data. This subset includes all 20 participants and 17 categorical-choice questions, yielding 340 participant–question pairs. Each pair is evaluated under both the Core-10 and Full-interview conditions, resulting in 680 condition-specific reasoning traces.

Human annotators were graduate students with backgrounds in NLP and/or psychology and were trained using the annotation guidelines described above. For each reasoning trace, the annotation schema records the model's predicted answer, supporting evidence excerpt, evidence location, and reasoning category.

Annotation Scope. Unlike our preliminary analysis, which used a stratified sample of users and questions, the final reasoning audit covers all categorical-choice items in the evaluation set. This design avoids selectively sampling only questions with large condition differences and provides a more comprehensive view of how models use interview-derived evidence across decision scenarios.

LLM-Assisted Pre-Annotation. For each instance, a large language model first produces structured outputs including the reasoning explanation, evidence excerpt, evidence location, and reasoning category. Human annotators then verify, correct, or refine these pre-annotated labels rather than labeling from scratch, focusing effort on error detection, ambiguous evidence boundaries, and cases where the model's explanation does not clearly support its predicted answer.

Human Verification and Reliability. To assess annotation reliability, we use the manually verified subset from the preliminary reasoning audit. This verification set contains 120 unique condition-specific reasoning traces: 60 overlap traces independently verified by all three annotators for inter-annotator agreement, and 60 additional coverage traces verified by a single annotator. This design yields 240 human annotation assignments in total. The resulting inter-rater agreement among human annotators reaches 95%, while agreement between human verification and the LLM pre-labels is 87.9%. These results support the use of the full pre-labeled categorical-choice dataset for large-scale descriptive reasoning analysis, while retaining human oversight for boundary cases and label consistency.

D.2 Reasoning Type definitions

To support the reasoning analysis, we manually defined a lightweight coding scheme as table D1 for model reasoning signals. The reasoning categories were designed to capture whether model decisions were grounded in specific episodes, abstracted participant values, or practical constraints. Despite the prompting protocol explicitly discouraging generic norm-based reasoning, there are still very few cases appeared in the final annotated dataset.

D.3 Evidence Location and Prediction Accuracy

D.3.1 Reasoning Type Shifts

Constraint-based reasoning more frequently incorporates follow-up-derived evidence than value-based reasoning. In the Full interview condition, 57.9% of constraint-based reasoning traces involve follow-up information (either exclusively or jointly with core interview responses), compared with 39.4% of value-based reasoning traces. See Table D2 for details.

Table D1: Reasoning signal categories used in the reasoning analysis.

Category	**Definition**	**Illustrative example**
Narrative reference	Reasoning explicitly references a specific personal experience, story, or situation described in the interview.	*Interview:* "When I was switching jobs last year, I chose stability over passion because I had rent pressure." *Reasoning:* "Given that you previously prioritized financial stability during a job transition, you would likely choose option B."
Value abstraction	Reasoning infers a generalized value, preference, or stable priority from one or more personal statements rather than citing a single concrete event.	*Interview:* "I dislike confrontations." "I often let things go in relationships." "I prefer harmony over winning arguments." *Reasoning:* "You generally prioritize interpersonal harmony over asserting your own position."
Coping / constraint	Reasoning invokes practical constraints, trade-offs, or coping strategies that limit the enactment of stated values. This category captures compromise under real-world conditions rather than abstract preference alone.	*Interview:* "Ideally I value autonomy, but right now I need income stability." "I don't like my manager, but I stay because changing jobs is risky." *Reasoning:* "Although you value autonomy, the financial risk makes you more likely to choose the safer option."
Generic norm / fallback	Reasoning relies primarily on general social norms, default assumptions, or common-sense preferences without clear participant-specific grounding. This category captures cases where the model produces a plausible answer but does not connect it to interview evidence.	*Interview:* No clear evidence about how the participant would allocate shared money. *Reasoning:* "In the absence of a strong stated preference, splitting the money evenly is the simplest and fairest option."

Note. Generic norm / fallback reasoning occurs rarely in the final annotated dataset and is treated as a descriptive category rather than a participant-grounded reasoning strategy. We retain it in the coding scheme to distinguish genuinely evidence-grounded reasoning from cases where the model falls back on general norms or default assumptions.

Table D2: Distribution of evidence sources across reasoning categories in the Full interview condition on the 340-item categorical-choice reasoning subset. Values indicate the proportion of reasoning traces within each category assigned to each evidence source.

Reasoning Type	n	Both	Core Interview	Follow-up	Unclassified
Constraint-based	19	0.421	0.421	0.158	0.000
Value-based	312	0.308	0.603	0.087	0.003
Narrative	5	0.000	1.000	0.000	0.000
Generic Norm	4	0.000	0.000	0.000	1.000

D.3.2 Paired Transition Results

Table D3: Paired correctness transitions among Full-condition cases that incorporate follow-up-derived evidence.

Transition type	Count	Percentage
Unchanged wrong	67	50.0%
Unchanged correct	46	34.3%
Improved	15	11.2%
Worsened	6	4.5%

Among the 134 Full-condition cases that incorporated follow-up-derived evidence, 15 improved over the Core-10 prediction while 6 worsened. The remaining cases preserved the same correctness status, with 67 remaining incorrect and 46 remaining correct. See Table D3 for details.

D.4 Confidence Intervals

For the reasoning analysis (n = 340), to assess the robustness of our reported accuracies, we compute 95% confidence intervals using bootstrap resampling over individual prediction instances.

The confidence intervals are wider, reflecting higher variance in reasoning behavior. However, the key trends—such as improved accuracy when predictions are grounded in follow-up-derived evidence—remain consistent across bootstrap samples.

For completeness, we also report accuracy ag-

gregated at the condition level within the reasoning analysis subset: Full Interview achieves 0.424 [0.371, 0.476], while Core-10 achieves 0.415 [0.362, 0.471]. These intervals overlap substantially, consistent with the observation that additional context alone does not uniformly improve prediction accuracy.

Table D4: Accuracy by evidence grounding with 95% confidence intervals.

Condition	Accuracy	95% CI
Follow-up grounded	0.455	[0.373, 0.545]
Not grounded	0.393	[0.328, 0.463]