Dialogue as Elicitation: Building High-Fidelity Personas for Personalized AI Advisors with LLMs

Weishi Shao
Independent Researcher
1983sirsws@gmail.com

Abstract

Dialogue as Elicitation (DaE) is a framework for building high-fidelity, reusable user personas that enable personalized AI advisors powered by large language models (LLMs). Instead of passively inferring preferences from sparse interaction logs, DaE treats persona construction as an explicit, structured elicitation process. We formalize a PersonaProfile schema spanning background, capabilities, resources, values, constraints, goals, weaknesses, challenges, and strategic paths. Operationally, DaE uses a two-agent architecture—an Elicitation Agent and an Application Agent—and a four-stage pipeline: initial collection, contradiction analysis with Socratic questioning, iterative refinement, and strategic pathfinding. We provide a minimal, model-agnostic prompt that implements this process with mainstream LLMs. In a small user study (N=5), participants rated DaE personas highly on coverage and accuracy (three out of five participants gave 5/5 coverage; four out of five gave 5/5 accuracy), and all preferred DaE personas over baseline self-written profiles for initializing a long-term advisor. These results suggest that structured persona elicitation can reduce alignment debt and improve immediate personalization from the first turn. We situate DaE within preference elicitation, user modeling, and multi-agent LLM coaching, and outline directions for scalable evaluation and integration in real-world advisory systems.

Keywords: LLMs; personalization; persona elicitation; preference elicitation; user modeling; multi-agent systems; Socratic questioning; prompt engineering; AI advisors; human-computer interaction

1 Introduction

Personalized large language models (LLMs) and "personal AI advisors" are rapidly proliferating. From career planning and long-term learning to health management and financial planning, many users hope that "from the very first turn, the AI gives advice as if it already knows me well." In practice, the experience is often the opposite:

- when a new chat is created, the model knows almost nothing about the user and can only start from generic suggestions;
- systems slowly and noisily piece together a "user profile" from scattered dialogues, logs, and behavioral traces;
- users struggle to fill in complex questionnaires or write a systematic self-introduction in one go.

The core bottleneck is thus not model intelligence, but the difficulty of constructing a high-quality user persona.

Existing personalization approaches roughly fall into three categories. First, *passive inference from behavioral data*: recommender systems, personalized learning, and user modeling traditionally infer preferences and abilities from clicks, browsing, or answers to exercises [1, 2]. Second, *fixed persona descriptions for dialogue models*: many works predefine persona snippets for fictional characters or NPCs and train models to act consistently with these scripts [3,4]. Third, *lightweight system prompts*: a small number of self-reported attributes or a short profile is embedded in the system prompt as a cheap persona.

These paths are effective in practice but share two limitations. (1) Persona construction is treated as *pre-work* or *hidden labor*: data collection, cleaning, and integration are often left to engineers and omitted from system papers [5]. (2) There is no explicit interaction paradigm dedicated to *persona elicitation*: dialogues are mostly designed to solve tasks, not to systematically surface and refine a user's values, constraints, and long-term goals.

In this paper we adopt a different perspective:

> For high-level personalized advisors, dialogue itself can be designed as an *active persona elicitation process—Dialogue as Elicitation (DaE).*

We treat an intensive, structured conversation between a user and an agent as a collaborative process for constructing a reusable *personal strategic profile* (`PersonaProfile`). This profile can then serve as shared system context for multiple *application agents* (e.g., career advisor, learning coach, decision assistant), allowing them to act in alignment with the user's background, resources, goals, and implicit constraints from the first interaction.

Contributions. Building on this view, we make three concrete contributions:

1. **Paradigm and problem definition.** We propose the DaE paradigm, framing high-quality persona construction as an active elicitation problem centered on Socratic dialogue rather than a passive by-product of data collection. We formalize the structure of `PersonaProfile` and discuss intuitive quality dimensions such as coverage, accuracy, and alignment.

2. **Method and architecture.** We present a dual-agent architecture that separates an elicitation agent from application agents, and a four-stage pipeline for persona elicitation: initial collection, contradiction analysis, iterative refinement, and strategic pathfinding. We provide a minimal deployable prompt that enables mainstream LLMs to implement DaE.

3. **Preliminary evaluation.** In a preliminary study with five real users, we compare baseline self-written personas against DaE personas in terms of subjective experience. Participants rate coverage and accuracy on a 1–5 Likert scale and choose which persona they would prefer to hand to a personal AI advisor. Results show that three participants assign a coverage score of 5/5 and two 4/5; four assign an accuracy score of 5/5 and one 4/5; all five prefer the DaE persona as advisor context.

Although the sample size is small, these results provide preliminary evidence that DaE can improve perceived persona quality and usefulness, and they offer a reusable template for more systematic evaluations.

2 Related Work

This work relates to preference elicitation, user modeling and personalization, LLM coaching and multi-agent orchestration, prompt engineering patterns, and emerging discussions of hidden labor in AI alignment.

2.1 Preference Elicitation

Classical preference elicitation treats user preferences as a latent utility function approximated through structured queries (e.g., option selection or pairwise comparisons) [6]. Recent work extends elicitation toward explainable step-wise reasoning in logic puzzles [7]. In contrast, DaE targets a *narrative personal strategic profile*: the output is a readable, discussable, and reusable document rather than an optimized numeric function. We borrow the interactive spirit of preference elicitation—iterative questioning and clarification—without adopting its numeric modeling objectives.

2.2 User Modeling and Personalization

Recent applied work in personalization emphasizes constructing user models to improve learning and creative support, e.g., topic-sensitive personalized learning [1] and creative writing preferences [2]. Historical dialog-centric user modeling frameworks anticipated explicit, inspectable user models [8]. These efforts confirm that high-quality user models can meaningfully improve system performance. However, construction is usually treated as an internal engineering artifact using implicit behavioral signals. DaE instead exposes persona construction at the dialogue layer: the model collaborates with the user to build an *explicit* `PersonaProfile` rather than relying solely on implicit traces.

2.3 LLM Coaching, Reflection, and Multi-Agent Paradigms

Recent studies explore LLMs as learning coaches, reflection assistants, or components of multi-role systems [3, 4, 9]. Structured multi-turn prompting and role orchestration (e.g., "teacher–student–critic") can improve quality and user experience. DaE differs by elevating persona elicitation itself as the *primary product*, decoupled from downstream tasks: application agents consume the elicited profile, while the elicitation agent specializes in high-fidelity persona construction.

2.4 Hidden Labor and Alignment Costs

Work on alignment and interaction cost [5] highlights the invisible labor users perform—repeatedly re-explaining background, correcting misunderstandings, and providing clarifications. Complementary perspectives on incentive alignment in human–AI collaboration [10] suggest structured workflows can improve outcomes when users invest focused effort. DaE concentrates this labor into a bounded, high-density structured dialogue, producing a reusable `PersonaProfile`. Rather than assuming users will gradually provide all needed details, DaE designs an explicit elicitation ritual with clear completion criteria and portability across advisors.

3 Method: Dialogue as Elicitation (DaE)

We formalize DaE through a problem definition, a `PersonaProfile` structure, a dual-agent architecture, and a four-stage elicitation pipeline. Figures 1 and 2 illustrate the architecture and interaction flow.

3.1 Problem Definition and PersonaProfile Structure

We consider users who desire a persistent "personal AI advisor" that supports decision-making, planning, and reflection over months or years. The persona must (i) have a long life cycle beyond a single session, (ii) be reusable across multiple application domains (career, learning, life planning), and (iii) balance specificity with openness for future updates.

The `PersonaProfile` is a multi-field structured document including (but not limited to): Background; Capabilities; Resources; Drives / Values; Constraints; Goals; Weaknesses & Patterns; Challenges & Uncertainties; Strategic Paths. We emphasize three intuitive quality dimensions: *coverage* (proportion of important user facts captured), *accuracy/fidelity* (alignment with the user's self-perception), and *alignment* (degree to which advisor outputs honor long-term values and constraints). Table 1 summarizes the fields.

3.2 Dual-Agent Architecture

DaE separates (1) an *elicitation agent* from (2) *application agents*. The elicitation agent acts as a strategic and cognitive coach, performing structured questioning, contradiction surfacing, and iterative rewriting. It adopts a professional, direct, "devil's advocate" style aimed at depth over comfort. Application agents (career coach, learning coach, planner) use the finalized profile to ground recommendations, highlight conflicts between short-term gains and long-term goals, and request clarification only when necessary. This separation allows different underlying models or deployment surfaces, and conceptually elevates persona construction to an independent interaction phase.

Field	Brief description
Background	Key background, life trajectory, important relationships
Capabilities	Core abilities and sustainable skills
Resources	Accessible external resources and channels
Drives / Values	Core drives and value priorities
Constraints	Time, financial, health, family constraints
Goals	Mid-term and long-term goals
Weaknesses & Patterns	Typical weaknesses and behavioral patterns
Challenges & Uncertainties	Current challenges and key uncertainties
Strategic Paths	Candidate strategic paths (conservative / balanced / aggressive)

Table 1: PersonaProfile structure overview.

3.3 Four-Stage Elicitation Pipeline

The elicitation agent follows four stages (Figure 2):

1. **Initial Collection**: User provides an initial self-portrait via an 8-item scaffold (background, capabilities, resources, financial/lifestyle context, core drives, goals, weaknesses/patterns, current challenges/uncertainties).

2. **Contradiction Analysis & Deep Questioning**: Agent scans for ambiguous abstractions (e.g., "freedom"), narrative inconsistencies, and value conflicts (e.g., prioritizing absolute security while seeking extreme autonomy). It conducts focused Socratic micro-dialogues to concretize terms, surface underlying motivations, and force prioritizations.

3. **Iterative Refinement**: After each deep questioning thread, the agent proposes rewritten profile segments removing empty phrases, adding concrete, verifiable details, and clarifying value orderings and constraints.

4. **Strategic Pathfinding**: Once stable, the agent and user co-create candidate strategic paths (e.g., conservative / balanced / aggressive), analyzing resource utilization, risk/reward at different time horizons, and alignment with drives and constraints.

This process repeats until major abstractions are grounded and a `PersonaProfile` v1.x emerges.

3.4 Execution Strategy and Prompt Overview

We provide a minimal prompt (not reproduced in full for space) specifying: (i) a strict role (no hollow encouragement, prioritize incisive questions), (ii) explicit stage management (agent signals transitions), (iii) targeted questioning templates for common vague terms, and (iv) structured local output formats (Observation / Quote / Question / Rewrite Proposal) and global full-profile outputs. Prompt pattern catalogs can further systematize prompt design choices [11], and broader LLM personalization methods complement DaE's elicitation focus [12]. This enables mainstream LLMs to implement DaE without extensive fine-tuning.

4 Preliminary Evaluation

We address: (1) Do users perceive high coverage? (2) Do they perceive high accuracy/fidelity? (3) Do they prefer DaE personas as long-term advisor context?

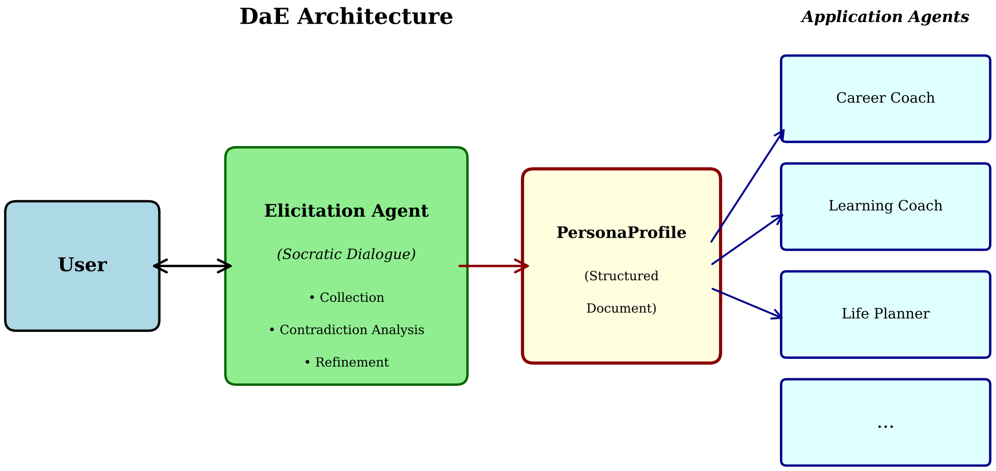

Figure 1: Overall architecture of Dialogue as Elicitation (DaE). The elicitation agent collaborates with the user to construct a reusable PersonaProfile document that can be shared across multiple application agents (e.g., career coach, learning coach, life planner).

4.1 Participants and Task Setup

Five adult participants (P1–P5) of varying familiarity with the author each produced two personas: a *baseline* self-written persona (answering open-ended prompts about background, current state, goals) and a *DaE* persona created via the four-stage elicitation flow. After completion, each participant rated the DaE persona (coverage, accuracy; 1–5 Likert) and chose which persona they would hand to a hypothetical long-term AI advisor.

4.2 Results Overview

Coverage: three participants rated 5/5, two 4/5 (none below 4). Accuracy: four rated 5/5, one 4/5 (none below 4). Preference: all five preferred the DaE persona over the baseline for advisor use. Table 2 summarizes ratings. These results suggest DaE produces personas perceived as both comprehensive and faithful.

4.3 Qualitative Observations

Interviews indicate participants felt DaE personas were more systematic, especially regarding weaknesses, implicit constraints, and behavioral patterns. Some noted mild psychological pressure from Socratic questioning but credited it with surfacing contradictions and deeper motivations. A typical sentiment: "If I have to give an AI one version, DaE captures my real conflicts and long-term aims better."

Limitations. The study is exploratory: N=5 is insufficient for fine-grained statistical analysis; social familiarity may introduce bias; metrics are subjective rather than downstream task performance. Despite these limits, consistent high ratings and unanimous preference provide initial evidence of perceived value and a template for future evaluations.

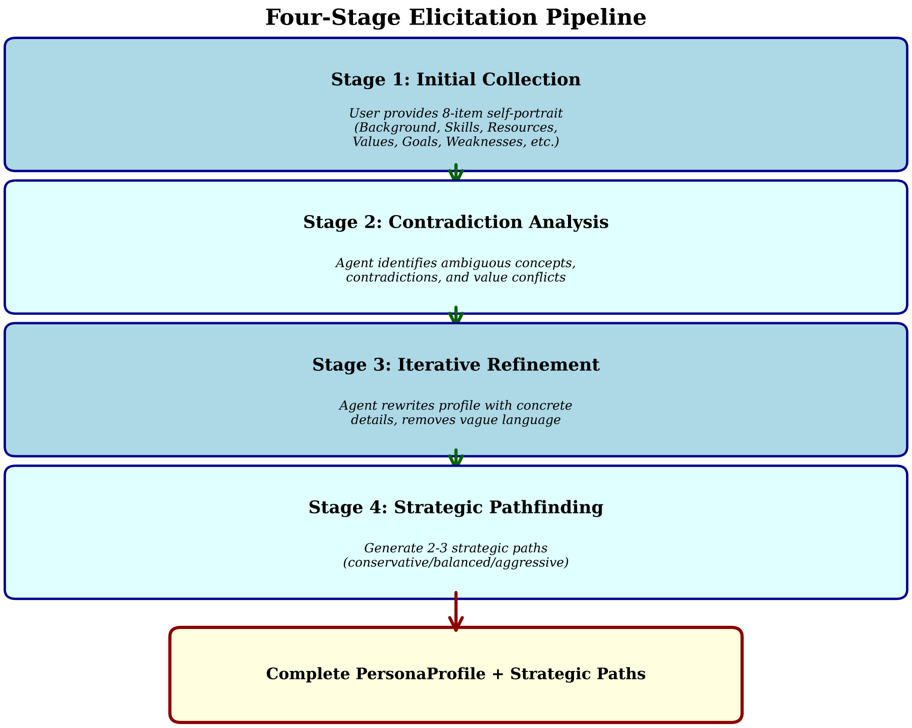
Four-Stage Elicitation Pipeline
Stage 1: Initial Collection
User provides 8-item self-portrait
(Background, Skills, Resources,
Values, Goals, Weaknesses, etc.)
Stage 2: Contradiction Analysis
Agent identifies ambiguous concepts,
contradictions, and value conflicts
Stage 3: Iterative Refinement
Agent rewrites profile with concrete
details, removes vague language
Stage 4: Strategic Pathfinding
Generate 2-3 strategic paths
(conservative/balanced/aggressive)
Complete PersonaProfile + Strategic Paths

Figure 2: Four-stage elicitation flow in DaE: initial collection, contradiction analysis and deep questioning, iterative refinement, and strategic pathfinding.

Dimension	5-point ratings	4-point ratings	≤3-point ratings
Coverage	3	2	0
Accuracy	4	1	0

Table 2: Summary of subjective ratings from 5 participants on coverage and accuracy of DaE personas. All participants preferred DaE personas over baseline self-written personas.

5 Discussion and Future Work

5.1 Ethical Considerations and Limitations

DaE explicitly engages with private, potentially sensitive information. Deployments should therefore adopt privacy-by-design principles: obtain informed consent, minimize data collection, encrypt storage, and provide granular access controls and deletion options. The elicitation flow can induce emotional strain; operators should set expectations, allow pausing, and avoid overreach into clinical territory.

This paper presents a small-N, exploratory study. Results reflect subjective judgments and potential selection bias. We did not quantify alignment beyond qualitative assessments; future work will operationalize and measure alignment, and compare outcomes across elicitation depths, user segments, and advisor tasks. Finally, even high-fidelity profiles risk staleness; systems should support profile versioning and revision to respect evolving goals and constraints.

5.2 Application Scenarios

DaE naturally suits long-term advisory settings: (i) career planning, entrepreneurship, life planning; (ii) learning coaching (customizing curricula and feedback using stable drives and constraints); (iii) selective health or reflective contexts (with strong privacy controls). In deployment, DaE can be an “episodic” module—users invest focused time once (or periodically) and multiple application agents reuse the profile.

5.3 Cost, Privacy, and Psychological Burden

Persona construction entails alignment labor: attention, time, and emotional effort. DaE acknowledges this explicitly, offering a bounded high-density ritual instead of diffuse, repeated clarification. Practical implementations must handle secure storage, granular access control, export/deletion rights, and pacing safeguards for emotionally charged questioning.

5.4 Future Work

Three directions: (1) Larger, more diverse human studies with controlled variations in questioning depth and style. (2) Measuring downstream advisor performance differences between baseline vs. DaE personas (e.g., career planning quality, user satisfaction). (3) Semi-automated persona quality checks (LLM-assisted consistency, abstraction grounding metrics, information extraction for coverage estimation). We hope the DaE framework and prompt design seed more systematic research on elicited persona quality.

6 Conclusion

We introduced Dialogue as Elicitation (DaE), treating high-quality persona construction as an active, structured dialogue rather than a passive artifact of data collection. We defined a reusable `PersonaProfile` structure, presented a dual-agent architecture and four-stage elicitation pipeline, and supplied a minimal prompt for mainstream LLM deployment. A small study (N=5) showed high subjective coverage and accuracy and unanimous preference for DaE personas over baseline self-written versions. We argue that the bottleneck for high-level personalized AI lies less in model capability and more in the absence of an honest, structured user strategic profile. DaE concentrates alignment labor into a portable elicitation ritual, moving toward AI advisors that genuinely act in harmony with long-term user values.

References

[1] Jieun Han, Daniel Lee, Haneul Yoo, Jinsung Yoon, Junyeong Park, et al. One-topic-doesn’t-fit-all: Transcreating reading comprehension test for personalized learning, 2025. arXiv:2511.09135.

[2] John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yi Wang, Yuqian Yuan, et al. Literarytaste: A preference dataset for creative writing personalization, 2025. arXiv:2511.09310.

[3] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *ACL*, 2016. Early work on persona-based dialogue generation.

[4] Martin Braas and Lukas Esterle. Fixed-persona slms with modular memory: Scalable npc dialogue on consumer hardware, 2025. arXiv:2511.10277.

[5] Cumi Oyemike, Elizabeth Akpan, and Pierre Hervé-Berdys. Alignment debt: The hidden work of making ai usable, 2025. arXiv:2511.09663.

[6] Paolo Viappiani, Boi Faltings, and Pearl Pu. Preference-based search using example-critiquing with suggestions. volume 27, pages 465–503, 2006. Interactive preference elicitation.

[7] Marco Foschini, Marianne Defresne, Emilio Gamba, Bart Bogaerts, and Tias Guns. Preference elicitation for step-wise explanations in logic puzzles, 2025. arXiv:2511.10436.

[8] Alfred Kobsa. Generic user modeling systems. In *User Modeling 2001*, pages 107–118. Springer, 2001.

[9] Neil K. R. Sehgal, Hita Kambhamettu, Sai Preethi Matam, Lyle Ungar, and Sharath Chandra Guntuku. Designing mental-health chatbots for indian adolescents: Mixed-methods evidence, a boundary-object lens, and a design-tensions framework, 2025. arXiv:2511.07729.

[10] Joshua Holstein, Patrick Hemmer, Gerhard Satzger, and Wei Sun. When thinking pays off: Incentive alignment for human-ai collaboration, 2025. arXiv:2511.09612.

[11] Jules White, Quchen Fu, Sam Hays, Peter Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv:2302.11382, 2023.

[12] Alireza Salemi, Mohammad Aliannejadi, Fabio Crestani, and W Bruce Croft. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023. arXiv:2304.11406.