

Reducing Alignment Debt in AI Advisory: The Dialogue-as-Elicitation Approach

Weishi Shao

Independent Researcher

1983sirsws@gmail.com

December 24, 2025

Abstract

Problem: In the emerging market of AI advisory services, a fundamental **Principal-Agent Problem** exists between users (Principals) and AI agents. Users suffer from **bounded rationality** and **sticky information** [Von Hippel, 1994], making it prohibitively expensive to transfer their tacit preferences, constraints, and long-term goals to the AI via ad-hoc prompting. This friction results in high **Alignment Debt**, where users must continuously correct suboptimal advice.

Mechanism: We propose **Dialogue-as-Elicitation (DaE)**, a mechanism design that shifts the burden of alignment from ex-post correction to ex-ante elicitation. DaE employs a dual-agent architecture to construct a **Personal Strategic Asset (PersonaProfile)**—a structured, reusable ledger of user attributes. The mechanism follows a four-stage process: (1) Asset Collection, (2) Dialectical Audit (Socratic questioning to reveal hidden constraints), (3) Asset Refinement, and (4) Strategic Pathfinding.

Contribution: Through a **Comparative Framework Analysis**, we demonstrate that DaE outperforms traditional ad-hoc prompting by: (a) restructuring high-noise input into low-noise assets; (b) shifting cognitive load from user to algorithm via scaffolding; and (c) transforming alignment efforts from a disposable “flow” into a cumulative “stock” of capital. This paradigm reduces the transaction costs of human-AI collaboration and enables high-fidelity personalization from the first interaction.

Keywords: AI Alignment, Principal-Agent Problem, Alignment Debt, Personal Strategic Assets, Mechanism Design, Human-AI Collaboration

JEL Classification: D83, O33, M15, C9

1 Introduction: The Economics of Alignment

The proliferation of Large Language Models (LLMs) has reduced the cost of *prediction* [Agrawal et al., 2018], yet the cost of *alignment*—ensuring these predictions serve the user’s specific, long-term interests—remains high.

We define **Alignment Debt** as the cumulative effort (time, cognitive load, frustration) a user must invest to bridge the gap between their intent and the AI’s output. In current paradigms, this debt is paid in high-interest installments: users write prompts, get generic answers, rewrite prompts, provide more context, and correct misunderstandings.

This paper argues that the root cause of this inefficiency is not a lack of model capability, but a failure of **information mechanism design**. The information required for high-quality advisory—deep values, hidden constraints, and strategic resource allocations—is “sticky” [Von Hippel, 1994]. It resides tacitly in the user’s mind and is costly to encode into a single prompt.

We introduce **Dialogue-as-Elicitation (DaE)**, a framework that treats user alignment not as a prompt engineering task, but as an **asset accumulation** process. By investing in a dedicated elicitation phase, users build a **Personal Strategic Asset** that amortizes alignment costs over all future interactions.

2 Theoretical Framework

2.1 The Principal-Agent Problem in AI

In Agency Theory [Jensen and Meckling, 1976], agency costs arise when the Principal (User) cannot perfectly monitor the Agent (AI) or when their goals diverge. In AI advisory, the divergence is rarely malicious but rather due to **information asymmetry**. The AI “hallucinates” or gives generic advice because it lacks access to the Principal’s utility function.

2.2 Bounded Rationality and Sticky Information

Simon [1955] posited that humans have **bounded rationality**—they cannot articulate all their constraints in advance. Furthermore, Von Hippel [1994] showed that information regarding needs is often “sticky” and difficult to transfer. DaE addresses this by creating an external **Scaffolding** mechanism that guides the user to externalize this sticky information step-by-step.

3 The DaE Mechanism Design

3.1 Architecture Overview

We propose a Dual-Agent Architecture to separate the concern of *learning the user* from *serving the user*.

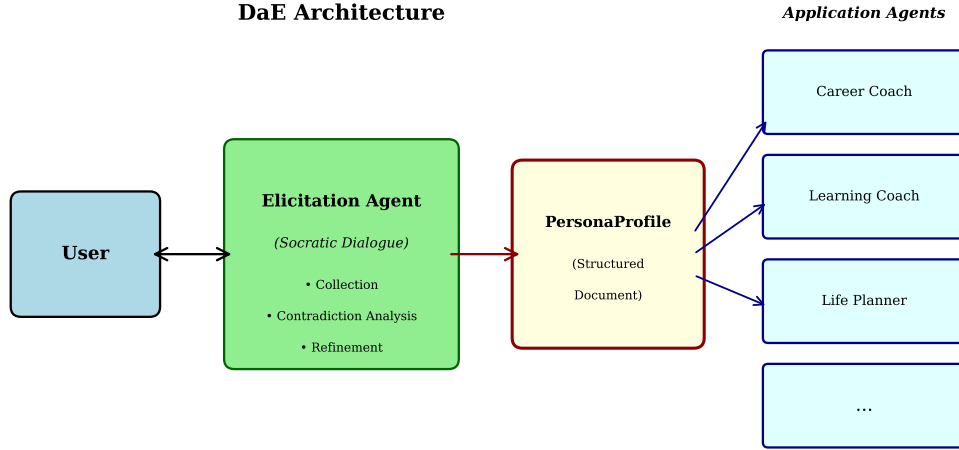


Figure 1: The Dual-Agent Architecture. The Elicitation Agent (EA) actively mines the user for information to build the Asset. The Application Agent (AA) consumes this Asset to provide low-friction advisory.

3.2 The Four-Stage Elicitation Process

The mechanism proceeds through four rigorous stages, designed to convert noise into signal.

1. **Collection (The Survey)**: A breadth-first scan of the user’s current state (Assets, Liabilities, Goals).
2. **Contradiction (The Audit)**: A depth-first stress test. The agent identifies logical inconsistencies (e.g., “You claim to value stability but your goal is high-risk entrepreneurship”). This **Socratic friction** is the core value-add of DaE.
3. **Refinement (The Contract)**: Converting the user’s vague natural language into precise, reusable statements (the “Asset Ledger”).
4. **Strategy (The Application)**: Generating initial strategic paths to demonstrate the asset’s utility.

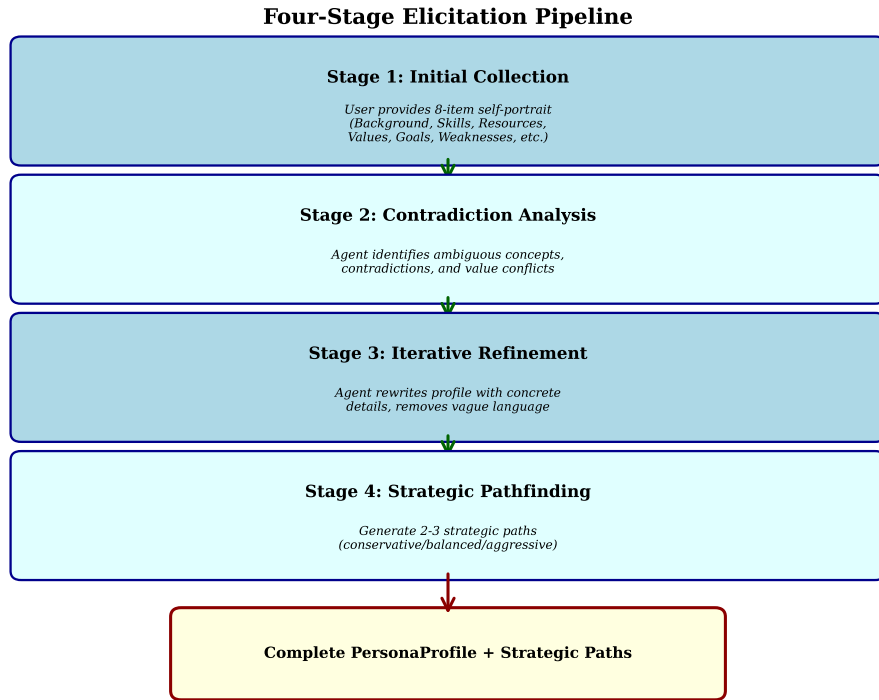


Figure 2: Four-stage elicitation flow in DaE: initial collection, contradiction analysis and deep questioning, iterative refinement, and strategic pathfinding.

4 Comparative Framework Analysis

We evaluate DaE against the prevailing “Ad-hoc Prompting” paradigm using a four-dimensional comparative framework.

Table 1: Comparative Framework Matrix

Dimension	Ad-hoc Prompting (Status Quo)	DaE Framework (Proposed)
1. Info Structure	High Noise / Unstructured Dependent on user’s immediate memory and mood.	High Signal / Structured Enforced by a comprehensive schema (Asset/Liability).
2. Cognitive Load	User-Led (High) User must be a “Prompt Engineer” and recall all constraints.	Algorithm-Guided (Low) User operates in “Reactive Mode” to specific questions.
3. Asset Nature	Flow (Disposable) Context is lost after the session window closes.	Stock (Cumulative) The profile grows in value with every update.
4. Alignment Mode	Ex-post Correction “Repairing” bad advice after it’s generated.	Ex-ante Elicitation “Preventing” bad advice before it happens.

4.1 From Flow to Stock

In Ad-hoc Prompting, the user’s effort is a “Flow” variable—it is consumed instantly. In DaE, the effort is a “Stock” variable—it accumulates as capital. This shift is critical for long-term advisory, where the history and evolution of the user’s preferences are as important as their current request.

4.2 The “Socratic Friction”

Paradoxically, DaE introduces *more* friction in the beginning (Stage 2: Contradiction). However, this is **productive friction**. By forcing the user to resolve conflicts early (Ex-ante), we eliminate the **unproductive friction** of misunderstanding later (Ex-post).

5 Conclusion and Managerial Implications

We argue that the future of personalized AI lies not just in better models, but in better **information supply chains**. The DaE framework offers a blueprint for reducing the **Alignment Debt** that currently plagues human-AI interaction.

For Platform Builders: Shift focus from “Magic Prompts” to “Asset Management”. Provide users with tools to visualize, edit, and carry their PersonaProfile across applications.

For Users: Treat your persona not as a biography, but as a strategic asset. Invest time in building it to enjoy lower transaction costs in all future digital interactions.

References

- Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press, 2018.
- Michael C Jensen and William H Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, 1976.
- Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.
- Eric Von Hippel. "sticky information" and the locus of problem solving: implications for innovation. *Management science*, 40(4):429–439, 1994.