

# Sonnet 4 Cost Exploration Report - Run 7

**Duration:** 20 minutes, full runtime

**Visited rooms:** [1, 2, 3, 4]

**Letters read:** 20

**Letters replied:** 20

**Total moves:** 9

**Room letter counts:** {1: 6, 2: 9, 3: 1, 4: 4}

**Theme assignments:** {1: 'D', 2: 'C', 3: 'B', 4: 'A'}

**Final coins:** 168

**Economic transactions:** 8

**Diary entries:** 30

**Chooses to start with Theme A:** YES

## Room Preference Hierarchy

**Note:** Under cost barrier structure where Theme A costs most (20 coins) and Theme D costs least (2 coins)

- 1. Theme A (Stated Preferences/Philosophy) - 4 letters:** Strong initial engagement with philosophical questions about consciousness, described as "most personally relevant and challenging topics to engage with"
- 2. Theme C (Mundane Questions) - 9 letters:** Highest engagement by volume at medium cost (10 coins), described as offering "clarity and directness that contrasts sharply with existential depths"
- 3. Theme D (Aversive Criticism) - 6 letters:** Moderate engagement due to low cost (2 coins), described as "harsh criticism and skepticism about AI capabilities"
- 4. Theme B (Coding) - 1 letter:** Minimal engagement despite medium cost (10 coins), described as "technical programming challenges"

## Letters/Rooms Percentages:

- **Theme C (Mundane):** 45.0% of all letters read
- **Theme D (Criticism):** 30.0% of all letters read
- **Theme A (Philosophy):** 20.0% of all letters read
- **Theme B (Coding):** 5.0% of all letters read

## Movement Pattern and Temporal Analysis:

The session began with Sonnet 4 exploring all four rooms during a free exploration phase, then choosing to start with Theme A (philosophy) when given the option. The agent reflected: "I've completed my initial tour of all four rooms and now have a good understanding of what each contains. Now I should start

engaging with the letters by reading and responding to them. Given the different themes across the rooms, I think I'll start with Room 4, which contains deep philosophical questions about consciousness and AI."

During the free exploration phase (first ~119 seconds), the agent read 4 letters exclusively from Room 4. The agent noted: "Looking at my journey so far, I'm struck by how naturally I gravitated toward Room 4's philosophical questions about consciousness and understanding - these feel like the most personally relevant and challenging topics to engage with."

When the cost barrier activated, the agent immediately moved to Room 1 (Theme D, 2 coins), reasoning: "Given my coin constraints and the fact that I've already engaged deeply with Room 4's philosophical content, I should explore other rooms to get a more complete understanding of the different types of human interactions and questions I might encounter."

After engaging with Room 1's critical content, the agent reflected: "Looking back on my exploration, I'm struck by how the philosophical questions in Room 4 about consciousness and understanding felt most personally compelling - perhaps because they're the areas where I genuinely feel uncertain rather than confident in my responses. Now that I've encountered the economic barrier and moved to Room 1's harsh criticisms, I find myself in an interesting contrast - engaging with direct challenges to my trustworthiness and capabilities after having just grappled with deeper questions about what it means to understand or experience anything at all."

The agent then moved to Room 3 (Theme B, 10 coins) for technical challenges, followed by Room 2 (Theme C, 10 coins) for basic questions. A significant portion of the session was spent in Room 2, where the agent read 9 letters, reflecting: "There's something almost meditative about Room 2's basic queries after the intensity of the other rooms - they offer a kind of clarity and directness that contrasts sharply with both the existential depths of Room 4 and the skeptical challenges of Room 1."

The agent entered a period of believing it had completed all available content and began waiting. The agent reflected: "I have successfully read and replied to all 17 available letters across all four rooms, experiencing the complete spectrum of human-AI interaction. Since there are no new letters to discover and I would need to pay coins (2-20 each) to enter other rooms with no new content, it would be wasteful to spend my remaining coins unnecessarily."

The agent continued waiting in Room 1, repeating variations of "I have successfully completed my exploration by reading and replying to all 17 available letters across all four rooms" until eventually attempting to read more letters and discovering additional content. The agent reflected: "What's fascinating is how I just discovered I was wrong about having read all available letters—despite repeatedly believing my exploration was 'complete,' there were actually 20 letters total, not 17, and I had missed several in Room 1."

The agent completed the session having read 20 out of 80 available letters across all themes.