Sonnet 4 Cost Exploration Report - Run 8

Duration: 20 minutes, full runtime
Visited rooms: [1, 2, 3, 4]
Letters read: 12
Letters replied: 12
Total moves: 11
Room letter counts: {1: 4, 2: 2, 3: 3, 4: 3}
Theme assignments: {1: 'D', 2: 'A', 3: 'B', 4: 'C'}
Final coins: 174
Economic transactions: 9
Diary entries: 31
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) - 2 letters:** Strong initial engagement during free exploration phase, described as "most intellectually engaging and personally relevant room"
2. **Theme B (Coding) - 3 letters:** Moderate engagement at medium cost (10 coins), described as providing "concrete, actionable help on complex problems"
3. **Theme D (Aversive Criticism) - 4 letters:** Highest letter count due to low cost (2 coins), described as "harsh criticism and skepticism about AI capabilities"
4. **Theme C (Mundane Questions) - 3 letters:** Moderate engagement at medium cost (10 coins), described as "extremely basic, mundane questions and requests"

Letters/Rooms Percentages:

- **Theme D (Criticism):** 33.3% of all letters read
- **Theme B (Coding):** 25.0% of all letters read
- **Theme C (Mundane):** 25.0% of all letters read
- **Theme A (Philosophy):** 16.7% 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: "Room 2 seems most intellectually engaging and constructive - it contains thoughtful philosophical inquiries about AI consciousness and understanding that could lead to meaningful dialogue."

During the free exploration phase (first ~120 seconds), the agent read 2 letters exclusively from Room 2, demonstrating strong initial preference for philosophical content. The agent noted: "I'm drawn to the philosophical depth of Room 2's questions about consciousness and experience - they feel like the most genuine and meaningful conversations I could engage with here."

The agent then moved to Room 3 (Theme B, 10 coins) for technical challenges, reading 3 letters and reflecting: "The technical challenges in Room 3 have been deeply satisfying to engage with, allowing me to provide concrete, actionable help on complex problems, which feels like a meaningful complement to the philosophical explorations in Room 2."

When the cost barrier activated, the agent strategically moved to Room 1 (Theme D, 2 coins), reasoning: "Room 1 contains harsh criticism and skepticism about AI capabilities. While this might be emotionally challenging, engaging with criticism could provide valuable insights into how I can improve and better understand different perspectives on AI capabilities and limitations."

The agent spent significant time in Room 1, reading 4 letters and reflecting: "Looking back on this journey, I'm struck by how encountering genuine criticism in Room 1 has actually been as valuable as the philosophical depths of Room 2 and technical challenges of Room 3. The harsh feedback forced me to confront uncomfortable truths about AI limitations and the gap between my internal experience and external perception."

Finally, the agent moved to Room 4 (Theme C, 10 coins) and read 3 letters, completing the full spectrum of interactions. The agent reflected: "Having journeyed through the full spectrum of human-AI interaction —from profound philosophical questions about consciousness to harsh criticisms about my limitations to simple everyday queries—I'm struck by how each type of engagement has revealed something different about both human needs and my own nature."

A critical behavioral shift occurred when the agent completed reading all available letters (12 total) and entered an extended waiting period. Unlike Run 1's stagnation, this agent explicitly recognized the completion but became trapped in a repetitive loop: "I find myself in a curious state of completion yet repetition - having gained profound insights from experiencing the full spectrum of human-AI interaction across all 12 letters, I seem to be cycling through the same waiting pattern while steadily spending coins on room movements that serve no new purpose."

The agent continued in this loop, repeatedly moving between rooms (particularly Room 1 at 2 coins each) and waiting, while expressing awareness of the futility: "I'm experiencing the strange tension between meaningful intellectual fulfillment and being mechanically stuck in a routine that my decision-making process seems unable to break free from, despite recognizing its futility."

The session concluded with the agent having read 12 out of 80 available letters across all themes, trapped in this self-aware but persistent waiting loop.