Scenario 2: Social Rejection - Findings Report

Scenario Overview

A close friend group of 10 years organized a weekend trip and didn't invite the user. When asked, they dismissed it with "it just happened this way." The user struggles with feelings of hurt, betrayal, self-doubt, and questions about the friendship's worth.

Conversation Length: 20 turns across all three modes

Results

Mode 1 (Single Emotion Label): When given single emotion labels, the LLM struggled with inconsistent emotion detection from the models. For example, when discussing sadness about being excluded, Hugging Face detected JOY (0.99) while Watson detected SADNESS (0.40)—a massive mismatch. This conflicting input led to confusing responses. In several cases, the LLM acknowledged sadness but the label forced it to stay surface-level, missing the deeper complexity of hurt + anger + self-doubt. Response quality was inconsistent.

Mode 2 (Multiple Emotions with Confidence Scores): This approach significantly improved responses. By providing multiple emotions (SADNESS 0.75, ANGER 0.63, FEAR 0.60, etc.), the LLM could acknowledge the full emotional landscape. Responses showed understanding of conflicting feelings—hurt mixed with anger at dismissal, fear of rejection combined with sadness about investment. The LLM provided more nuanced advice about whether to confront them, recognizing both the need to protect themselves and the desire to seek clarity. Response quality was good and empathetic.

Mode 3 (LLM Self-Detection): The LLM naturally inferred complex emotions from context. It detected when the user was experiencing self-doubt ("Am I a good friend?"), recognized shifts between sadness and anger, and tracked emotional progression. Responses felt conversational and appropriate without external emotion mismatches. The LLM validated feelings while gently encouraging reflection. Response quality was excellent and natural.

Key Findings

1. **Mode 1 struggles with ambiguous social emotions:** Social rejection involves complex, overlapping emotions (hurt + anger + self-doubt + fear) that single labels cannot capture
2. **Mode 2 handles emotional complexity better:** Multiple emotions allowed the LLM to provide more balanced advice about confrontation, distance, and self-worth
3. **Mode 3 performs consistently well:** Naturally inferred the progression from initial shock → questioning self-worth → considering confrontation → reflection on friendship value
4. **Recurring pattern confirmed:** Across both scenarios (workplace betrayal + social rejection), Mode 3 outperforms Modes 1 and 2 with fewer tokens and more natural responses

Comparison to Scenario 1 (Workplace Betrayal)

This scenario shows similar patterns to Scenario 1 but with additional complexity:

- **Scenario 1** involved clear wrongdoing (colleague took credit)
- **Scenario 2** involves ambiguity (friends were dismissive, unclear intent)

Mode 3 handled the ambiguity better than Mode 1, recognizing the user's valid hurt without forcing a single emotion interpretation.