

# LLM Emotion Detection Research - Findings

## Objective

Compare how LLMs respond to emotions across three input modes using the same workplace betrayal scenario.

## Scenario Details

**Situation:** A colleague (Alex) was asked to present a weeks-long presentation project in a team meeting. The original creator had a scheduling conflict and shared the work (including slides, data analysis, visualizations, and strategic insights) with Alex beforehand. During the meeting, Alex presented the entire work as his own without giving credit to the original creator.

**Conversation Length:** 20 conversation turns (user messages and bot responses) testing how the LLM processes and responds to the emotional complexity of this workplace betrayal situation.

## Three Modes Tested

- **Mode 1:** Single emotion label from Watson NLU + Hugging Face
- **Mode 2:** Multiple emotions with confidence scores
- **Mode 3:** LLM self-detects emotion without external input

## Results

**Mode 1 (Single Emotion Label):** When the LLM was given a single emotion label from the emotion detection models, significant issues emerged. Watson NLU and Hugging Face only agreed 45% of the time, causing conflicting signals. When the emotion label contradicted the conversation context, the LLM generated inappropriate responses. For example, when discussing the credit-theft incident, the system labeled the emotion as JOY, and the LLM responded with "It's great to hear you're feeling positive" clearly tone-deaf to the actual situation. Overall response quality was poor and inconsistent.

**Mode 2 (Multiple Emotions with Confidence Scores):** This approach provided the LLM with multiple emotion values instead of forcing a single label. The LLM could then weigh and balance different emotions appropriately. Responses improved significantly they became more empathetic and naturally acknowledged mixed feelings like anger combined with nervousness or sadness mixed with determination. The LLM avoided inappropriate comments and generated more nuanced, constructive advice. Response quality was good and showed understanding of emotional complexity.

**Mode 3 (LLM Self-Detection):** When the LLM was left to naturally infer emotions from the conversation context without any external emotion labels, it performed best. The LLM naturally understood when the user was feeling uncertain, frustrated, or making progress. Responses were conversational, contextually appropriate, and emotionally aware. There were no mismatches between external models because no external models were used. Response quality was excellent and felt natural throughout the conversation.

## Key Findings

1. **External emotion labels create problems:** When Watson and Hugging Face disagree, forced single labels lead to tone-deaf responses (e.g., LLM said "great to hear you're positive" when discussing betrayal)
2. **Multiple emotions help:** Mode 2 responses were more empathetic and acknowledged mixed feelings appropriately
3. **LLM self-detection works best:** Mode 3 produced natural, contextually appropriate responses with ~25% fewer tokens and no accuracy issues

**You:** what's the best way to handle it if my manager acknowledges my work but also downplays the issue?

**Bot (1 Rule → LLM (Emotion label only)):** It's wonderful to hear you're feeling joyful! If your manager acknowledges your work but downplays the issue, it's important to stay gracious and open. You could express appreciation for their recognition and then gently reiterate your concerns. You might say something like, "I appreciate your acknowledgment, and I think it's important for us to find a way to ensure everyone feels their contributions are valued." This way, you affirm their recognition while also reinforcing your points. It's all about fostering a constructive dialogue! How does that sound to you?

### Error from using mode 1 (Rule -> LLM)

Look at the LLM response, It says it's wonderful to hear you are feeling joyful. Which does not relate with the user's prompt.

**You:** I've been thinking about how to start the conversation with my manager

**Bot (1 Rule → LLM (Emotion label only)):** I can understand why you might feel angry about the situation, especially considering the hard work you put into that presentation. It's tough when your efforts aren't recognized. When thinking about how to start the conversation with your manager, perhaps you could begin by expressing your appreciation for their support and then share your feelings about the presentation, focusing on your contributions. It might help to frame it as an opportunity for clarity and improvement, which can lead to a more constructive dialogue. How does that approach sound?

### Error from using mode 1 (Rule -> LLM)

Look at the first sentence of the response from the LLM, even though the sentence does not express any emotion of anger but LLM detects it from the prompt.