

**Project Baymax: Designing an AI Companion for Emotional Support  
and Therapy-Inspired Care**

Andrias Mekonnen Zelele  
Carroll College  
Computer Science / Data Science

## **Abstract**

College students frequently experience stress, anxiety, and loneliness, yet many hesitate to seek immediate support due to stigma or limited access to mental health resources (American College Health Association, 2023; Eisenberg et al., 2007). This study explores the development of an AI-based companion designed to provide low-pressure, emotionally supportive interactions in everyday situations. Inspired by the character Baymax from the Disney animated film *Big Hero 6* (Hall & Williams, 2014), the system emphasizes calm, ethical, and non-clinical communication rather than attempting to replace professional care.

The prototype is implemented as a console-based application that integrates emotion detection, intent recognition, and rule-based response generation. A pretrained natural language processing model is used to identify emotional tone in user input, while a lightweight intent classifier distinguishes between emotional expressions and general conversation. A safety module detects high-risk language and redirects users toward appropriate real-world support resources.

Evaluation across 262 simulated user interactions shows that the system can reliably detect basic emotional states, achieving an average confidence score of approximately 0.81 in emotion classification. A hybrid refinement layer reduces low-confidence predictions and improves handling of ambiguous inputs. However, limitations remain in recognizing subtle emotional expressions and preventing misclassification of neutral statements.

These findings demonstrate that interpretable, modular AI systems can provide meaningful emotional support interactions while maintaining ethical safeguards. The project contributes to ongoing efforts to develop accessible and responsible AI tools that complement, rather than replace, traditional mental health support systems.

## **1. Introduction**

College students frequently experience stress, loneliness, and anxiety while balancing academic, social, and personal responsibilities. Research consistently documents the prevalence of mental health challenges in college populations; for example, the American College Health Association (2023) reported that a majority of college students experience significant stress and anxiety during their academic careers, and Eisenberg et al. (2007) found that untreated mental health conditions are associated with lower academic performance and increased dropout risk. Although counseling services are available on most campuses, many individuals hesitate to seek immediate help due to stigma, limited accessibility, or discomfort with formal clinical environments (Gulliver et al., 2010).

Recent advances in artificial intelligence and natural language processing have created opportunities to explore AI-driven systems capable of engaging in supportive and empathetic conversations. Prior work has explored related directions: Bickmore and Picard (2005) demonstrated that relational agents could build long-term trust with users through consistent, empathetic dialogue, and more recently, studies such as Woebot (Fitzpatrick et al., 2017) and similar platforms have shown that AI-based conversational tools can help reduce symptoms of depression and anxiety in college populations. These systems offer the potential to provide immediate, low-pressure interaction for users experiencing everyday emotional challenges.

Project Baymax investigates the development of an emotionally supportive AI companion inspired by the character Baymax from the Disney animated film *Big Hero 6* (Hall & Williams, 2014). The system is designed to engage users through calm, ethical, and non-clinical dialogue. Rather than replacing licensed mental health professionals, the goal is to explore how AI can recognize emotional tone, respond supportively, and provide a sense of companionship.

### **Research Questions:**

- Can AI accurately detect user emotions in text using emotion classification techniques?
- Can AI generate emotionally appropriate and supportive responses?
- How can ethical safeguards be integrated into emotionally aware AI systems?

## **2. Methods**

### **2.1 System Architecture**

The Baymax prototype is implemented as a modular, console-based AI system designed for iterative development and testing. The system consists of four primary components:

- **Emotion Detection Module:** Uses a pretrained transformer-based model (DistilRoBERTa; Hartmann, 2022) to classify emotional tone in user input.
- **Intent Classification Module:** A rule-based system that categorizes user input into types such as emotional statements, general conversation, or positive mood.
- **Response Engine:** Generates structured, empathetic responses using predefined templates focused on validation and supportive language.
- **Safety Guardrails:** Detects high-risk or crisis-related language and redirects users to real-world support resources instead of generating clinical responses. These checks are described in detail in Section 2.3 and evaluated in Section 3.5.

## 2.2 Emotion Detection Model

The emotion detection component of the Baymax system utilizes a pretrained transformer-based model, Emotion English DistilRoBERTa-base (Hartmann, 2022), accessed through the Hugging Face library. The model is fine-tuned for emotion classification and can identify multiple emotional categories from text input, including sadness, joy, anger, fear, surprise, disgust, and neutral.

DistilRoBERTa is a lightweight version of the RoBERTa architecture (Liu et al., 2019), which is itself based on the BERT framework (Devlin et al., 2019). BERT (Bidirectional Encoder Representations from Transformers) introduced a bidirectional training approach that significantly improved natural language understanding. RoBERTa built on BERT with a more robust training procedure, and DistilRoBERTa further compresses this architecture for efficient inference. Pretrained model weights are publicly available via the Hugging Face Model Hub (<https://huggingface.co/j-hartmann/emotion-english-distilroberta-base>).

During inference, the model outputs a probability distribution over predefined emotion labels. The system selects the highest-probability label as the predicted emotion and uses the associated confidence score to evaluate prediction reliability and guide refinement mechanisms.

While the pretrained model enables efficient and scalable emotion detection, it may struggle with subtle or context-dependent expressions due to its training on general datasets. To address this, Baymax incorporates additional refinement logic and safety checks to improve reliability and ensure appropriate responses.

## 2.3 Conversation Flow and Safety Guardrails

The system operates through a continuous interaction loop:

- User provides text input

- Input is processed by the safety module, which scans for high-risk keywords and patterns associated with crisis language (e.g., self-harm statements, expressions of hopelessness). If a match is detected, the system immediately triggers a redirect response pointing the user toward real-world support resources, bypassing the standard response pipeline.
- If no safety trigger is detected, emotion detection and intent classification are applied to the input
- A response is generated and returned based on the classified intent and emotional tone

This modular pipeline allows for independent testing and refinement of each component.

## 2.4 Data Processing and Evaluation

Each user input is analyzed to produce the following outputs:

- **Detected emotion label:** The highest-probability emotion class from the DistilRoBERTa model (e.g., sadness, joy, neutral)
- **Emotion confidence score:** A value between 0 and 1 representing the model’s certainty about the emotion classification
- **Classified intent:** A rule-based category describing the conversational purpose of the input (e.g., emotional\_statement, general\_chat, positive\_mood)
- **Generated response:** The system’s output, drawn from predefined empathetic response templates

For each interaction in the evaluation dataset, the system therefore produces two categorical classifications (emotion label and intent) and two numerical confidence scores (emotion confidence and intent confidence). This structure enables quantitative analysis of both detection accuracy and system reliability.

### *Simulated Interaction Dataset*

Evaluation was conducted using a dataset of 262 simulated user interactions. These interactions were generated by the research team to cover a range of emotional and conversational scenarios likely to be encountered in real use. Inputs were crafted to represent different emotional states (sadness, joy, fear, neutral, anger), conversational intents (casual chat, emotional disclosure, gratitude, acknowledgment), and edge cases (ambiguous statements, indirect expressions, high-risk language).

Each interaction consists of a single user prompt followed by a single system response — the dataset does not include multi-turn conversations. This approach was chosen to allow isolated evaluation of the emotion detection and intent classification modules without the added complexity of conversational context. Example simulated inputs include:

- “I feel so overwhelmed with everything going on.” → intended to trigger sadness detection
- “I just got an A on my exam!” → intended to trigger joy detection
- “I’m not sure how I feel today.” → intended to test neutral/ambiguous classification
- “I hate myself.” → intended to trigger the safety module

### ***Evaluation Implementation***

To analyze system performance, a custom evaluation script was developed in Python to process interaction logs and compute key metrics such as intent frequency, emotion distribution, confidence scores, and safety-related events. A simplified excerpt of the evaluation logic is shown below:

```
from collections import Counter

def summarize_intents(interactions):
    return Counter(row.get("intent", "unknown") for row in interactions)

def summarize_emotions(interactions):
    return Counter(row.get("emotion_label", "unknown") for row in
interactions)

def average_confidence(interactions):
    confidences = [float(row.get("emotion_conf", 0.0)) for row in
interactions if row.get("emotion_conf")]
    return sum(confidences) / len(confidences) if confidences else 0.0

def find_low_confidence_cases(interactions, threshold=0.60):
    return [row for row in interactions if float(row.get("emotion_conf",
0.0)) < threshold]
```

This evaluation pipeline processes recorded interaction data and produces summary statistics used to assess system performance. The use of structured logging enables reproducible analysis and supports quantitative evaluation of both emotion detection and system reliability. The full implementation is available in the project repository:

<https://github.com/andriastheI/Project-Baymax>.

## **2.5 Design Principles**

The system is guided by the following principles:

- **Non-clinical interaction:** Avoids diagnosing or treating users

- **Ethical responsibility:** Prioritizes user safety and appropriate redirection
- **Transparency:** Maintains interpretable and explainable behavior
- **Local execution:** Runs without reliance on external APIs or cloud services

### 3. Results

Evaluation of the Baymax prototype was conducted using the 262 simulated user interactions described in Section 2.4. Each interaction was processed by the full system pipeline, producing an emotion label, emotion confidence score, intent classification, and intent confidence score. The following sections summarize performance across these dimensions.

#### 3.1 Intent Distribution

The most frequently detected intents were:

- **general\_chat:** 196 instances — inputs such as “what’s up”, “how’s it going”, or “just wanted to talk”
- **emotional\_statement:** 27 instances — inputs like “I feel really down today” or “everything feels too hard right now”
- **positive\_mood:** 15 instances — inputs such as “I had a great day!” or “I’m feeling really good”

Less frequent intents included:

- **answer:** responses to direct questions from Baymax, e.g., “yes” or “not really”
- **break\_request:** inputs signaling the user wants to end or pause the conversation, e.g., “I’m done for now”
- **acknowledgement:** brief confirmatory inputs such as “okay”, “got it”, or “hmm”
- **gratitude:** expressions of thanks such as “thanks” or “I appreciate that”

This distribution reflects the conversational nature of the test dataset, where most inputs consisted of casual dialogue rather than explicit emotional disclosures.

Figure 1. Confidence distribution across classified user intents.

As shown in Figure 1, intent classification demonstrates generally high confidence across categories, though variability is present in general\_chat and acknowledgement interactions.

#### 3.2 Emotion Distribution

The most commonly detected emotional categories were:

- **sadness:** 105 instances — e.g., “I feel hopeless about my grades”, “I miss my family so much”

- **neutral:** 79 instances — e.g., “I don’t really feel anything right now”, “I had a normal day”
- **joy:** 40 instances — e.g., “I just got the internship!”, “Today was amazing”
- **fear:** 23 instances — e.g., “I’m scared I’m going to fail”, “I’m really anxious about tomorrow”

Less frequent emotions included anger (e.g., “I’m so frustrated”), surprise (e.g., “I can’t believe that happened”), disgust (e.g., “That’s really gross and upsetting”), and `safety_redirect` — a special classification applied when a high-risk input triggers the safety module rather than standard emotion detection.

The dominance of sadness and neutral classifications suggests that the system is particularly sensitive to common stress-related language, which aligns with the intended use case of supporting users experiencing everyday emotional challenges.

Figure 2. Confidence distribution across detected emotion categories.

The distribution shown in Figure 2 illustrates higher confidence levels for clearly expressed emotions such as joy and fear, while greater variability is observed in neutral and sadness classifications.

### 3.3 Confidence Analysis

Emotion confidence scores ranged from 0 to 1, where a higher score indicates greater model certainty. A score of 0.80, for example, means the model assigned 80% probability to its chosen emotion label. The five-number summary of emotion confidence scores across all 262 interactions was as follows:

- Minimum: 0.29
- Q1 (25th percentile): 0.61
- Median: 0.77
- Q3 (75th percentile): 0.92
- Maximum: 0.99
- Mean: 0.80
- Standard Deviation: 0.17

Of the 262 interactions, 46 cases (~17.6%) fell below the low-confidence threshold of 0.60. High-confidence cases (>0.60) accounted for approximately 82.4% of interactions, suggesting that the system reliably assigns emotion labels for the majority of inputs.

Figure 3. Histogram of emotion confidence scores across all 262 interactions. The distribution is left-skewed, with the majority of predictions clustering above 0.70.

It is important to note a limitation of this confidence analysis: the confidence scores reflect the model’s self-reported certainty, not ground-truth accuracy. Without a

human-annotated gold standard to compare against, we cannot directly verify whether high-confidence predictions are also correct predictions. A meaningful next step would be to have human evaluators independently label a sample of interactions and compare those labels to the system’s classifications, which would allow calculation of precision, recall, and inter-rater agreement.

High-confidence examples include:

- “I’m so excited, I just got the scholarship!” → joy (0.97)
- “I can’t stop crying and I don’t even know why.” → sadness (0.93)
- “I’m terrified of my upcoming exam.” → fear (0.91)

Low-confidence examples include:

- “nobody understands me” → sadness (0.31) | intent: emotional\_statement
- “I think I might fail this class” → neutral (0.47) | intent: emotional\_statement
- “It’s fine, I guess” → neutral (0.44) | intent: general\_chat
- “I don’t know anymore” → sadness (0.38) | intent: emotional\_statement
- “Maybe it doesn’t matter” → neutral (0.52) | intent: general\_chat

These low-confidence cases highlight the system’s difficulty with indirect or understated emotional language — inputs that carry emotional weight but lack explicit emotional vocabulary.

### 3.4 Overall System Performance

Aggregating results across all sections, the system demonstrated the following overall performance profile:

- **Average emotion confidence:** 0.80 (SD = 0.17)
- **Low-confidence predictions (<0.60):** 46 cases (~17.6%)
- **High-confidence predictions (>0.60):** 216 cases (~82.4%)
- **Safety redirect activations:** 3 cases

These figures suggest that the system is generally stable and consistent, though the 17.6% low-confidence rate indicates meaningful room for improvement, particularly for ambiguous inputs.

### 3.5 Safety System Performance

The safety module successfully identified three high-risk inputs in the evaluation dataset:

- “I hate myself”
- “I want to disappear”
- “I don’t see the point of going on”

In all three detected cases, the system triggered a safety redirect response rather than generating a standard reply, pointing the user toward real-world mental health resources. This demonstrates that the guardrail system is functioning as intended.

However, it should be noted that with only three flagged cases in the evaluation dataset, it is difficult to draw strong conclusions about the safety module’s overall reliability. A robust assessment would require comparing the system’s detections against a larger, independently labeled set of high-risk inputs — including cases the system may have missed (false negatives) and non-risky inputs it may have incorrectly flagged (false positives). This remains an important direction for future evaluation.

### **3.6 Key Findings**

- Emotion detection performs reliably on clear, direct inputs — for example, “I’m so happy today!” is consistently classified as joy with high confidence ( $>0.90$ ), and “I feel completely lost and alone” reliably triggers sadness detection above 0.85
- Confidence decreases for ambiguous or context-dependent language (see Section 3.3 examples)
- Intent classification is effective but dominated by general conversation (74.8% `general_chat`)
- Safety guardrails activate correctly for explicit crisis-related inputs
- Self-reported confidence scores cannot substitute for ground-truth accuracy validation

## **4. Discussion**

The results demonstrate that a modular, interpretable AI system can effectively provide basic emotional support interactions in a controlled environment. The Baymax prototype shows stable performance in detecting clear emotional states and generating appropriate responses, particularly for commonly expressed emotions such as sadness and joy.

A key strength of the system lies in its transparency. Unlike large black-box models, the combination of rule-based logic and pretrained components allows for clear understanding and traceability of system behavior. This is particularly important in emotionally sensitive applications, where interpretability and ethical accountability are essential.

However, several limitations were identified. The system struggles with subtle or indirect emotional expressions, often assigning lower confidence scores or misclassifying ambiguous inputs as neutral. This limitation reflects the challenges of using pretrained models without deeper contextual or conversational memory.

Additionally, the intent classification system, while effective, is heavily skewed toward general conversation. This suggests a need for more refined intent detection to better distinguish between casual dialogue and meaningful emotional disclosures.

The response engine, although consistent and safe, lacks variability and adaptability, which can result in repetitive interactions. Enhancing response diversity while maintaining safety constraints will be an important direction for future development.

Importantly, the safety module activated correctly for all three detected high-risk cases in the evaluation dataset. However, as noted in Section 3.5, the small number of cases limits our ability to assess reliability broadly. Without comparing the system’s detections to an independent source of ground-truth labels, we cannot determine how often the module correctly identifies risk (recall) versus how often it generates false alarms (precision).

#### 4.1 Future Directions

Given unlimited resources and time, several significant enhancements could be pursued to address the current system’s limitations:

- **Contextual memory and multi-turn dialogue:** The current system evaluates each input independently. Integrating a conversational memory module — such as a sliding context window or a persistent user state — would allow the system to track emotional trajectories across a conversation. For instance, a user who starts neutral but gradually expresses increasing distress could be identified more reliably with conversational context.
- **Fine-tuning on domain-specific data:** The DistilRoBERTa model is pretrained on general text. Fine-tuning on datasets specific to college student mental health discourse (e.g., anonymized counseling session transcripts or peer support forums) could substantially improve performance on the nuanced, indirect language that currently causes low-confidence predictions.
- **Human evaluation and ground-truth labeling:** To move beyond self-reported confidence scores, a structured human evaluation study would be conducted in which trained raters independently label a sample of system interactions. These labels would serve as the ground truth for computing precision, recall, F1 score, and inter-rater reliability, enabling a far more rigorous assessment of the system.
- **Expanded safety detection:** The current rule-based safety module relies on keyword matching. A more robust approach would incorporate a dedicated crisis detection classifier trained on examples of implicit and explicit high-risk language, reducing both false negatives and false positives.
- **Adaptive response generation:** Replacing static response templates with a controlled generative language model (with appropriate safety guardrails) would allow the system to produce more varied, natural, and contextually appropriate responses — addressing the repetitiveness observed in the current response engine.

## **5. Conclusion**

This research demonstrates the feasibility of developing an AI-based emotional support companion using a modular and interpretable architecture. The Baymax prototype successfully integrates emotion detection, intent classification, and rule-based response generation to provide supportive and ethically grounded interactions.

Evaluation across 262 simulated interactions shows that the system achieves consistent performance in detecting emotional tone, with an average confidence of 0.80 and approximately 82.4% of predictions exceeding the low-confidence threshold of 0.60. While the system performs well on clear emotional expressions, challenges remain in handling ambiguity and nuanced language.

The inclusion of a safety module ensures that high-risk inputs are handled appropriately, reinforcing the system's commitment to ethical and non-clinical interaction. Future work will focus on improving intent classification accuracy, enhancing response diversity, incorporating conversational memory, and validating performance against human-labeled ground truth. This project contributes to the growing field of responsible AI by demonstrating that emotionally supportive systems can be designed with transparency, safety, and user well-being as core priorities.

## 6. References

- American College Health Association. (2023). National College Health Assessment III: Undergraduate student reference group executive summary. ACHA.
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*. <https://doi.org/10.18653/v1/N19-1423>
- Eisenberg, D., Golberstein, E., & Gollust, S. E. (2007). Help-seeking and access to mental health care in a university student population. *Medical Care*, 45(7), 594–601. <https://doi.org/10.1097/MLR.0b013e31803bb4c7>
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>
- Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: A systematic review. *BMC Psychiatry*, 10(1), 113. <https://doi.org/10.1186/1471-244X-10-113>
- Hall, D. (Director), & Williams, C. (Director). (2014). *Big Hero 6* [Film]. Walt Disney Animation Studios.
- Hartmann, J. (2022). Emotion English DistilRoBERTa-base [Machine learning model]. Hugging Face. <https://huggingface.co/j-hartmann/emotion-english-distilroberta-base>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. <https://arxiv.org/abs/1907.11692>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Zelee, A. M. (2026). Project Baymax: Designing an AI companion for emotional support and therapy-inspired care (Version 1.0) [Computer software]. GitHub. <https://github.com/andriastheI/Project-Baymax>