CHULA INTERNATIONAL SCHOOL OF ΣNGINEERING
The ingenuity of CHULA ΣNGINEERING

Progress Report 1

Pookie: An AI-Driven Robot for Promoting Mental Wellbeing and Emotional Support

Authors: Tibet Buramarn, Kridbhume Chammanard, and Thitaya Divari

Advisor: Dr. Paulo Fernando Rocha Garcia and
Ms. Kunpariya Siripanit

2147416 Final Project I
International School of Engineering (ISE)
Chulalongkorn University

September 27, 2024

Contents

1 Introduction

This progress report provides an update on the development of Pookie, an AI-driven robot designed to promote mental well-being, developed in assistance from Chula Student Wellness. Pookie aims to act as a companion to help alleviate feelings of stress and anxiety, especially in response to future societal concerns like "Terror Outbursts," an anxiety-driven phenomenon anticipated to affect Thailand. The project focuses on enhancing positivity and emotional attachment by creating a robot that interacts with users in an empathetic and calming manner.

2 Overview

				MVP 1 - Proof of Concept & Customer-Centric Specifications															
Task Number	Task Title	POC	PCT OF TASK COMPLETE	September				October				November				December			
				Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1	**MVP 1 - Sprint 0**																		
1.1	Advisor Onboarding	KC	100%																
1.2	Project Scoping	KC,Nile,Kitty	100%																
1.3	Project Scoping - Expert Interviews	KC,Nile,Kitty	100%																
1.4	Project Proposal	KC,Nile,Kitty	100%																
1.5	In-Depth Technology Specifications	KC,Nile	100%																
1.6	E2E Customer Journey	KC,Kitty	75%																
2	**MVP 1 - Sprint 1**																		
2.0	Facial Expression Model - Data Collection	KC,Nile	75%																
2.1	Facial Expression Model - Training/Testing	KC,Nile	0%																
2.2	Facial Expression Model - Quality Testing	KC,Nile	0%																
2.3	SER Model - Data Collection	KC,Nile	100%																
2.4	SER Model - Training/Testing	KC,Nile	75%																
2.5	SER Model - Quality Testing	KC,Nile	0%																
2.6	UX Design (Fusion360)	KC,Kitty	25%																
3	**MVP 1 - Sprint 2**																		
3.1	Parts Procurement	KC,Kitty	0%																
3.2	Jetson Orin/Nano Test	KC,Nile,Kitty	0%																
3.3	MVP 1 Integration	KC,Nile,Kitty	0%																
3.4	Market Validation	KC	0%																
3.5	Final Presentation Preparations	KC,Nile,Kitty	0%																

Figure 1: Project GANTT Chart

Overall, the project progress has taken the correct trajectory ever since the project proposal was submitted. According to the project GANTT Chart shown in Figure 1, key objectives of Q3 and Q4 (17th September to 27th September, respectively) include in-depth technology specifications and an end-to-end customer journey. In-depth technology specifications refers to the specific software approach, particularly related to machine learning, that will yield the most effective yet feasible results. On the other hand, the end-to-end customer journey refers to a flow chart for each specific interaction between human and robot, specifying the trigger points and response for the robot in order to successfully capture the objective of the project: to build a robot that can understand human emotions on a basic level, and provide positive reinforcement to alleviate signs of stress and anxiety.

3 Facial Expression Recognition

The development of the Pookie AI-driven robot incorporates a crucial element: detecting and interpreting user emotions, particularly stress and anxiety, using Facial Emotion Recognition (FER). This section outlines the current progress, challenges, and future approaches for designing the emotion detection system, including relevant datasets, models, and methodologies.

3.1 Design Challenges

Designing the input-output interaction for the robot's FER systems presents significant challenges. Specifically:

- **Defining input/output structure:** A clear operational structure needs to be established for when the robot actively listens for inputs and how it produces outputs. The temporal window during which the robot processes emotional input and when it offers feedback is still under discussion for most effective interaction.
- **Universality of emotions:** Stress and anxiety, being complex and abstract emotional states, are not universally exhibited across all racial or cultural groups, complicating the design of an accurate detection system, particularly for Thai users.
- **Dataset Challenges:** A significant hurdle in developing the FER system is the lack of available facial expression datasets for Thai people or Asians more broadly. In Thailand, researchers from Mahidol have even claimed that an open-source dataset for facial expressions in Thai ethnicity does not conventionally exist for analysts and physicians. Since accurate emotion recognition relies heavily on data representative of the target user group, this absence poses a major limitation.

Given that the most concerning challenge is related to datasets, the team is in conversation with researchers from Mahidol University to acquire a dataset from their research: **MU Face Emotion - Building a Large Dataset for Emotional Facial Expression in Psychological Domain. However, the team has not received a response**[1]. As a result, the team is investigating research papers to identify datasets with facial expressions most similar to those of Thai individuals. This approach would involve leveraging such datasets, followed by transfer learning techniques to fine-tune a pre-trained model on the most culturally appropriate data available. One such open source dataset is **A Chinese Face Dataset with Dynamic Expressions and Diverse Ages Synthesized by Deep Learning**[2], where a team of researchers created a facial expression image generation model for various Chinese faces belonging to different age groups, genders, and face structure, as shown in Figure 2.

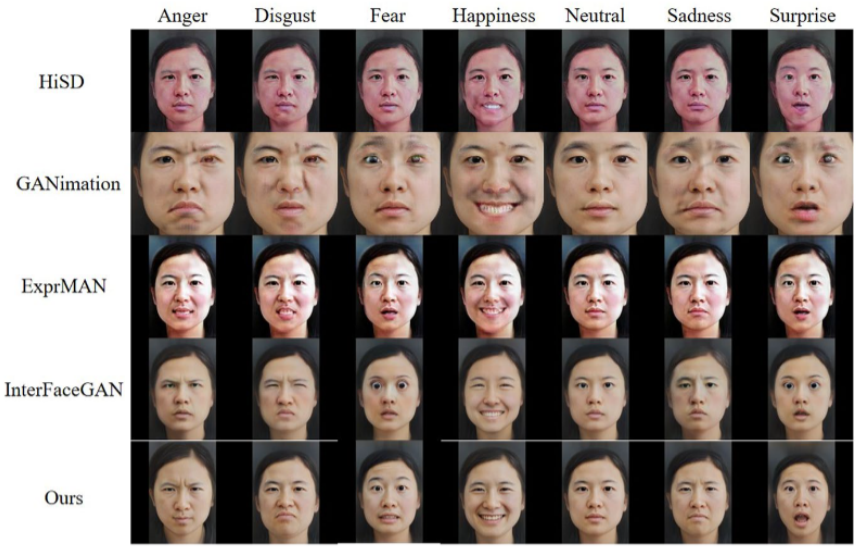

Figure 2: A Chinese Face Dataset with Dynamic Expressions and Diverse Ages Synthesized by Deep Learning ("Ours" represents the research's proposed model)

3.2 Facial Expression Models and Stress Detection

Facial emotion recognition is commonly based on the detection of six universal emotions: anger, disgust, fear, happiness, sadness, and surprise. These are well-researched and exhibit consistency across different cultures, making them a reliable basis for detecting emotions related to stress and anxiety. Stress and anxiety often manifest through combinations of universal emotions. For example, stress may be reflected through anger, disgust, or fear, while anxiety may be linked to fear or sadness. Since models for detecting stress and anxiety directly are difficult to build, particularly for Asian populations, the proposed method is to detect these emotional cues by focusing on anger, disgust, and fear[3]. These emotions have been scientifically linked to stress, as indicated by several studies.

3.3 Approach for FER in Pookie

1. The FER system for Pookie will use a two-pronged approach:
    - **Anxiety Detection via Speech Emotion Recognition:** Proven to be a reliable method, Pookie will detect anxiety by analyzing voice inputs, which may offer better accuracy.
    - **Stress Detection via Facial Emotion Recognition:** Stress will be detected primarily using facial expressions linked to anger, disgust, and fear, as these are the most indicative emotions for stress according to research.
2. Data Acquisition and Fine Tuning
    - Select a pre-trained model (e.g., VGG19 or OpenFace) that has been trained on a universal dataset.
    - Fine-tune the model using transfer learning by adding a new classifier layer, specifically trained on the newly acquired dataset with similar characteristics to the target population.
    - Freeze the convolutional layers of the original model and train higher layers on the Thai dataset, ensuring it can adapt to more culturally specific facial expressions.
3. Program Design
    - A stress classification function that monitors emotions such as anger, disgust, and fear within a defined time window.
    - The system evaluates whether these emotions surpass a threshold, indicating stress.

This progress marks an important step towards enhancing Pookie's emotional intelligence and its ability to support users' mental well-being.

4 Speech Emotion Recognition

Another key element of Pookie's AI is the Speech Emotion Recognition (SER) system, which is used for recognizing the user's emotions based on vocal patterns. By analyzing factors such as pitch, tone, and intensity, the system detects emotions: neutral, anger, happiness, sadness, and frustration. This section outlines the current progress, challenges, and future approaches for designing the speech emotion recognition model, including relevant datasets and methodologies.

4.1 Speech Emotion Recognition Model

We are utilizing a dataset created by Chulalongkorn University in collaboration with VISTEC, DEPA, and AIS, containing 41 hours and 36 minutes of audio recordings labeled with five emotions: neutral, anger, happiness, sadness, and frustration. VISTEC has also developed a speech emotion recognition model based on this dataset. We are in the process of adjusting the model's parameters to better fit our system.

Parameters:

- **Number of Mel-filterbanks:** Mel-filterbanks are used to transform the frequency spectrum into a scale that better aligns with how humans perceive sound. By adjusting the number of filterbanks, we can control the resolution of this transformation, which affects the granularity of frequency representation in the model. More filter banks provide finer detail, while fewer filterbanks reduce the model's sensitivity to frequency variations.
- **Sampling Rate:** The sampling rate refers to how many samples per second the audio is recorded or processed. A higher sampling rate captures more detail from the audio signal, but it also increases the computational load. Adjusting this parameter ensures that the audio quality is sufficient for emotion recognition without overwhelming system resources.
- **Frame Length of STFT:** The STFT converts audio signals into a time-frequency representation. The frame length defines how long each segment of audio is for this transformation. A longer frame provides more frequency resolution but less time precision, while a shorter frame offers better time resolution but less frequency detail. Balancing these is key for accurate emotion recognition.
- **Epochs:** This parameter refers to the number of times the entire training dataset passes through the model during training. Adjusting the number of epochs affects how well the model learns from the data. Too few epochs may lead to underfitting, where the model doesn't learn enough, while too many can cause overfitting, where the model memorizes the training data but performs poorly on new data.

The results are as shown in Figure 3, Figure 4, and Figure 5.

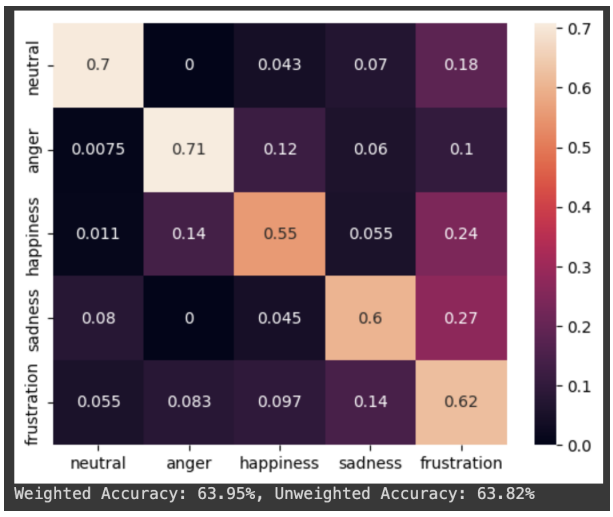

Figure 3: Default parameters with 80 mel-filterbanks, sampling rate at 16,000 Hertz, and frame length of STFT at 50 milliseconds at 80 epochs

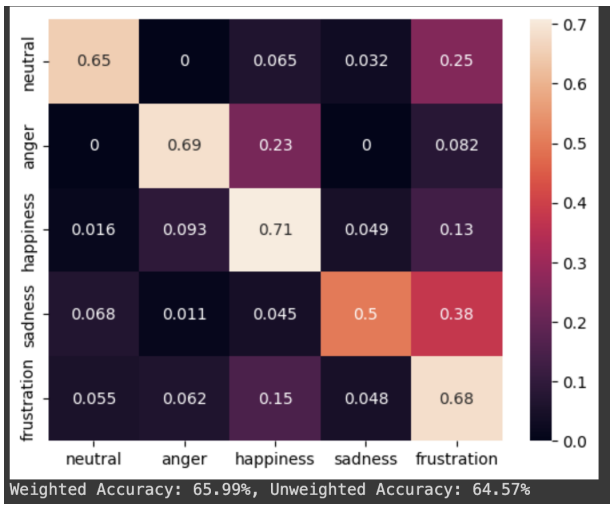

Figure 4: Tuned model with 128 mel-filterbanks, sampling rate at 22,050 Hertz, and frame length of STFT at 50 milliseconds at 60 epochs

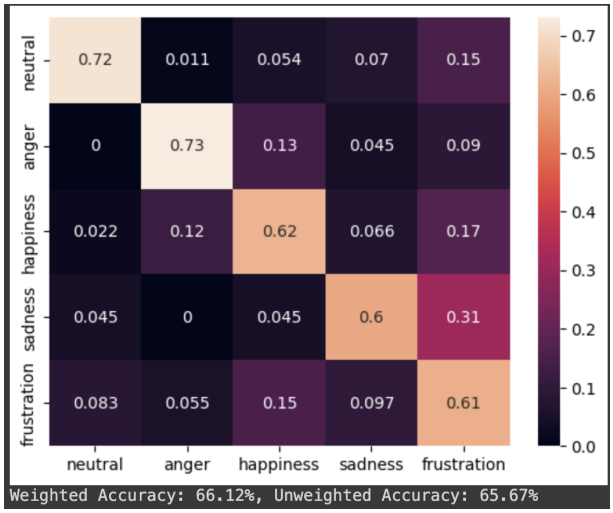

Figure 5: Tuned model with 128 mel-filterbanks, sampling rate at 16,000 Hertz, and frame length of STFT at 25 milliseconds at 60 epochs

4.2 Design Challenges

Similarly to the FER, the input-output interaction for SER also poses a significant challenge:

- **Defining input/output structure:** This design challenge which was addressed for FER is also present in SER as these two systems are intertwined and SER's initialization might depend on other factors such as duration of facial detection. As of the report this is still under discussion as mentioned in the FER section.

Overall, the current design issues of the SER stems from undecided input and output structure and interactions, both of which will be addressed by the next progress report in October.

4.3 Technical Challenges

Several technical challenges have emerged, primarily related to compatibility:

- **Compatibility issues:** The original SER model was created three years ago, and since then, Google Colab has updated its Python version from 3.7 to 3.10, alongside a CUDA update to version 12.1. Python 3.7 reached its end of life (EOL) as of June 27, 2023, which has caused compatibility issues with the model's required libraries. The mismatch between Google Colab's Python environment and the model's requirements has led to several errors. To address this, we created a forked version of the original VISTEC-SER model and modified it according to the updated dependencies which include a plethora of changes to the source code.

5 Robot Design

5.1 Component Identification and Selection

A critical aspect of the progress has been the careful selection of internal components that will drive the core functionality of the robot. After thorough research and evaluation, the following essential components have been identified:

- **NVIDIA Jetson Orin Nano (x1):** Serves as the primary processing unit
- **Speaker (x1):** Enables the robot to deliver audio feedback and engage in voice-based communication with users.
- **Pressure Sensors (x2):** Integrated for tactile responsiveness, allowing the robot to detect touch and adjust interactions based on user input.
- **Microphones (x2):** Facilitate audio input for natural language processing and communication.
- **Servo Motors (x5):** Power the robot's neck and arm movement.
- **Camera (x1):** Provides visual input.

5.2 Design Development

The cornerstone of progress has been the development of the robot's overall design. The concept revolves around an anthropomorphic robot inspired by a red panda, aimed at creating a visually appealing and emotionally engaging design.

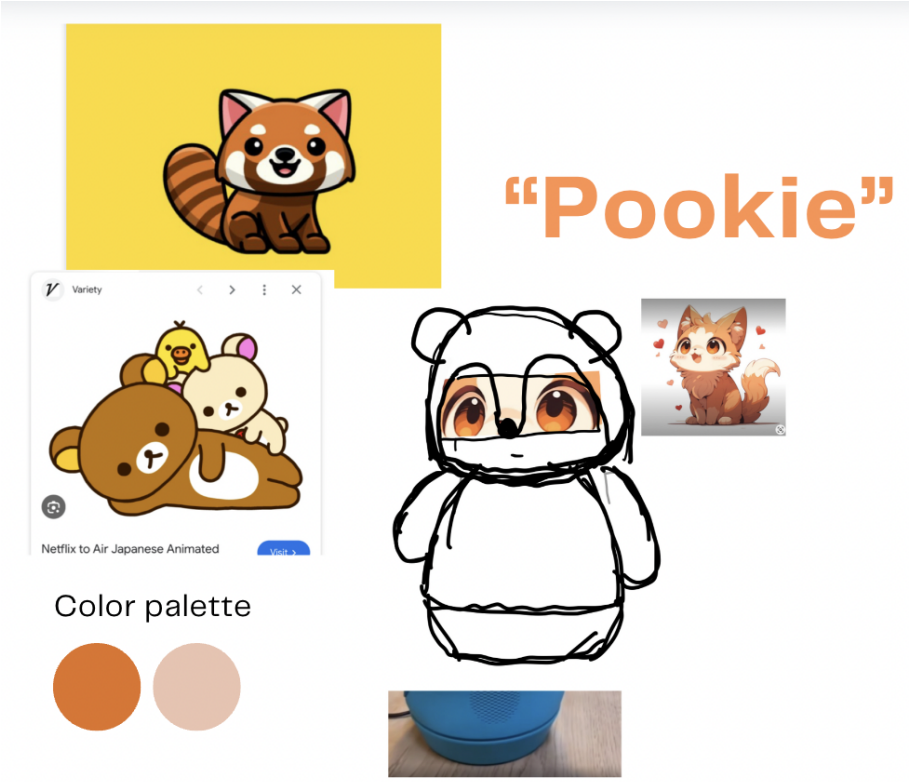

Figure 6: Pookie Anthropomorphic Design

Fusion 360 was utilized to visualize and refine the concept, creating a preliminary sketch of the robot. This 3D modeling software facilitated the determination of precise dimensions and allowed for quick iterations on different design elements. Various body proportions were explored, with features adjusted to achieve a balance between animal-like charm and human-like functionality.

The design process involved crafting an external shell while simultaneously considering the internal layout. This approach ensured that all previously selected components could be accommodated within the shell while maintaining the desired aesthetics.

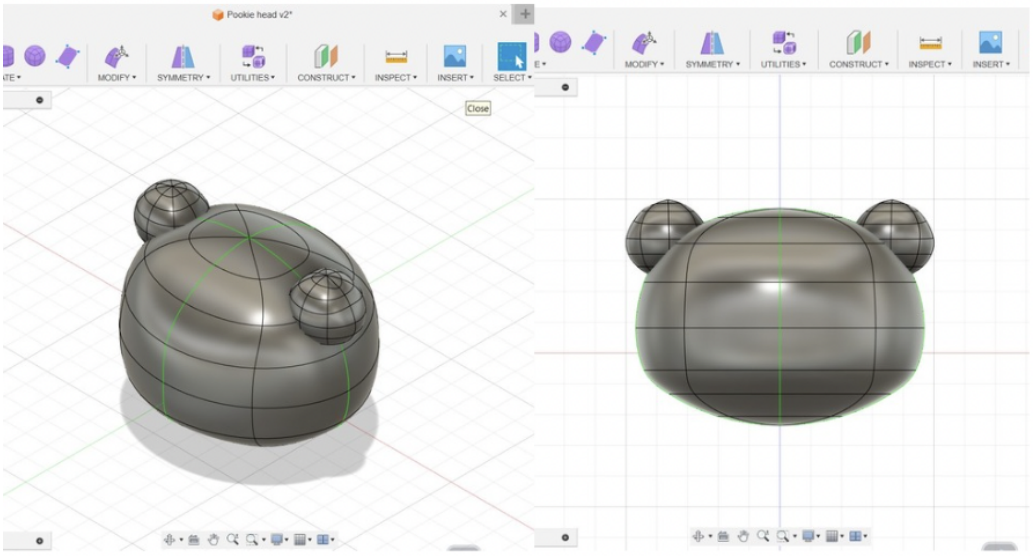

Figure 7: Pookie's Head Shell Design in Fusion360

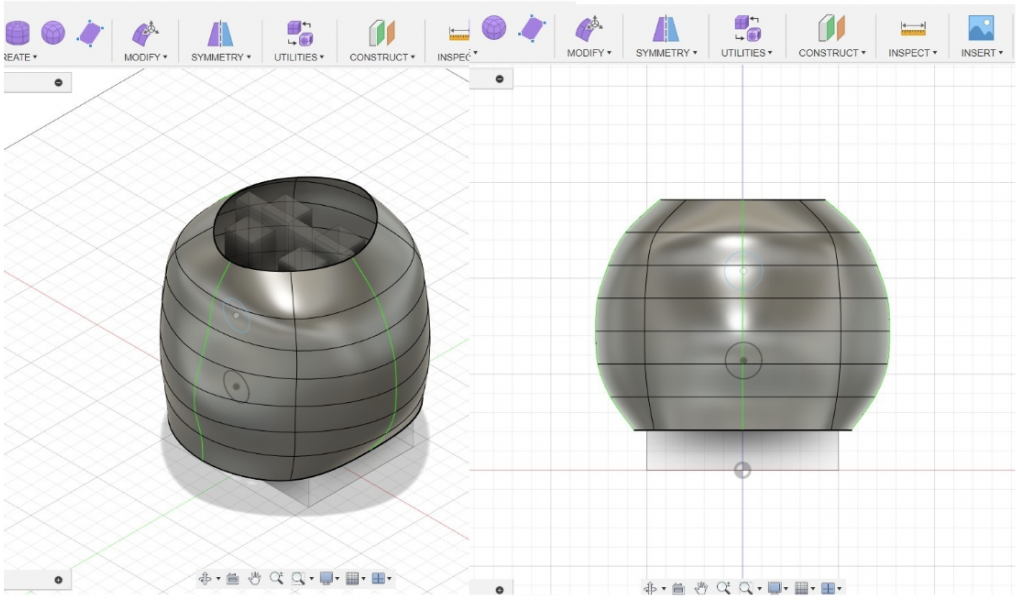

Figure 8: Pookie's Body Shell Design in Fusion360

Careful consideration was given to the robot's dimensions, aiming for a scale that would make it approachable yet large enough to house all necessary components and facilitate meaningful interactions. This iterative process resulted in a comprehensive 3D model that serves as an initial blueprint for Pookie the robot.

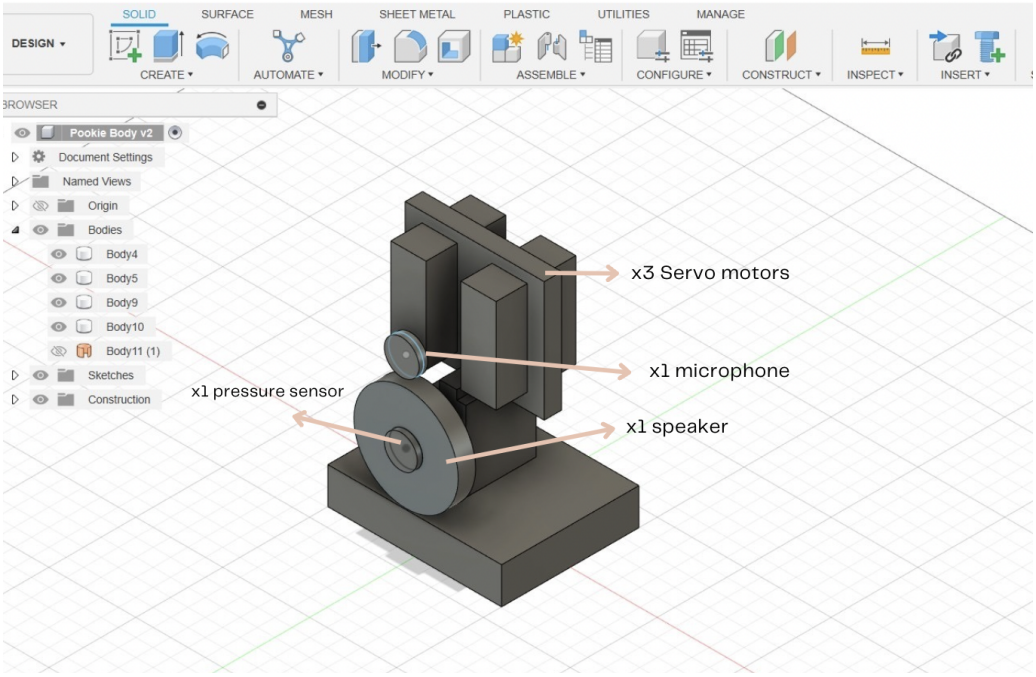

Figure 9: Component Dimensions Inside Pookie's Body Shell

5.3 Eyes Design

The robot's eye design is crucial for conveying emotion and engaging users. After reviewing numerous examples, optimal designs were selected based on their potential to humanize the robot and facilitate non-verbal communication. The selected LED eye designs are expected to integrate seamlessly with other expressive features, contributing significantly to creating a cohesive and engaging interaction experience.

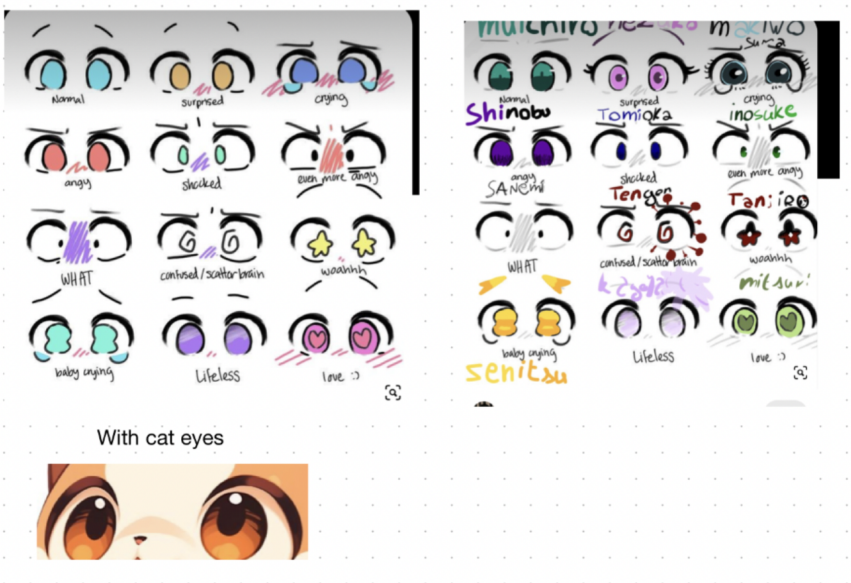

Figure 10: Pookie's LED Eye Design

5.4 Emotion Expression through Movement

The project has progressed to developing expressive movement patterns, enabling the robot to convey a spectrum of emotions through physical gestures. This phase involves mapping specific combinations of head tilts, body postures, and limb movements to effectively communicate various emotional states such as happiness, curiosity, confusion, and alertness.

References

[1] S. Jaidee, K. Wongpatikaseree, N. Hnoohom, S. Yuenyong, and P. Yomaboot, "Mu face emotion - building a large dataset for emotional facial expression in psychological domain," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, 2022, pp. 1–6.

[2] S. Han, Y. Guo, X. Zhou, J. Huang, L. Shen, and Y. Luo, "A chinese face dataset with dynamic expressions and diverse ages synthesized by deep learning," *Scientific Data*, vol. 10, 12 2023.

[3] J. Almeida. and F. Rodrigues., "Facial expression recognition system for stress detection with deep learning," in *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, INSTICC. SciTePress, 2021, pp. 256–263.