Towards LLM-Integrated Storytelling Robot to Improve College Student Mental Health

UNIVERSITY of SOUTH FLORIDA

RARE LAB

Tam Do, Chi Vo, Duy Linh Ha, Felix Mendez, Suyog Bam and Zhao Han

Introduction & Background

My Mental Health Has Worsened Since I Have Been in School

- ■ Disagree
- ■ Neither Agree or Disagree
- ■ Neither Agree or Disagree

| All Students (n=1,000) | 24% | 25% | 51% |

How Would You Rate Your Current Mental Health?

- ■ Excellent
- ■ Very Good
- ■ Good
- ■ Fair
- ■ Poor

| All Students (n=1,000) | 6% | 32% | 33% | 13% |

From College Student Mental Health Report by Best College in 2022

Why Robots? Why Storytelling?
Researchers have explored robotic interventions for well-being—but most lack emotionally rich, adaptive storytelling. Storytelling fosters empathy, emotional reflection, and social connection—key elements of mental health support.

We introduce a novel robot powered by a Large Language Model (LLM) that listens, generates dynamic stories, and expresses matching emotions through facial expressions and gestures. Unlike static systems, our robot responds in real-time to user input.
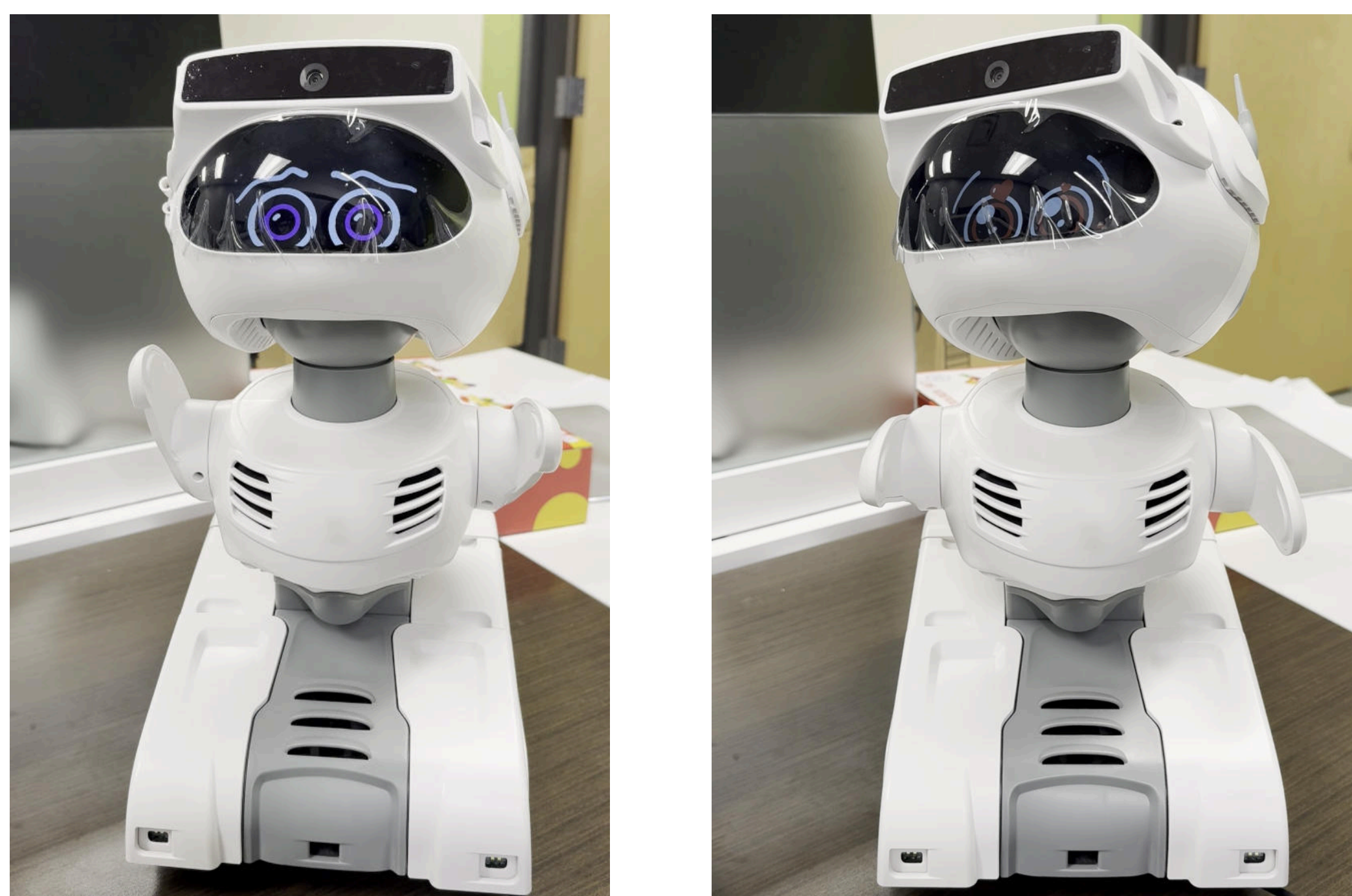
System Overview

Fig. 1. As preliminary results, we implemented a storytelling robot, Misty, using LLMs to dynamically generate stories with cohesive facial expressions, arm gestures, and head movements. Left: A loving face showed before saying "… looked at the rabbit and the berries [food]," following "We didn't have any food!" Right: An empathetic face showed at the beginning of the story before saying "but she had been trapped by the responsibilities for her family…" following "Alora always dreamed of traveling the world."
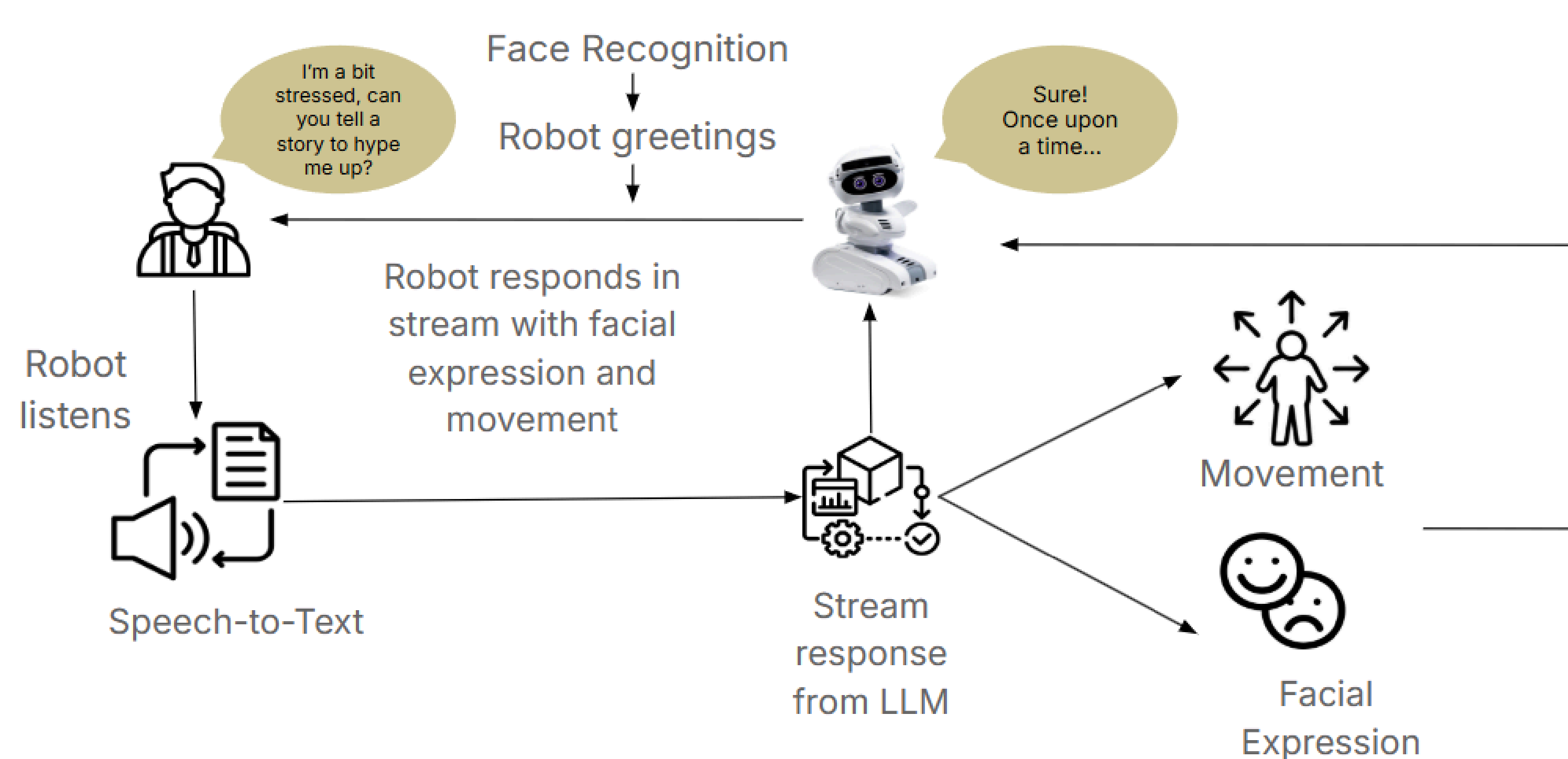

Fig. 2. The system workflow with LLMs to dynamically generate stories and corresponding facial expressions. We enhance voice expressiveness by replacing Misty's built-in TTS with Google TTS, making the robot's voice more human-like. Face detection triggers storytelling, with Misty responding to user inputs, displaying facial expressions, and cohesive arm and head movements corresponding to emotions at the sentence level.

Evaluation

We compared four most popular LLMs: Gemini 2.0 Flash, GPT-3.5, Llama 2, and DeepSeek R1 to choose the top performing model for our robot's storytelling ability.

Table 1: Evaluation of four LLMs based on storytelling relevance, emotional alignment, obedience to prompt constraints, naturalness, content structure, and latency.

Model	Relevance	Consisten.	Obedience	Natural.	Content & Len.	Latency
Gemini 2.0 Flash	High	High	High	High	Well-balanced	Fast (∼7.6s)
GPT-3.5	High	High	High	Moderate	Sometimes verbose	Moderate (12.6s)
Llama 2	Moderate	Low	Moderate	Low	Inconsistent	Moderate (11.7s)
DeepSeek R1	High	High	High	High	Detailed but long	Slow (∼19.2s)

Table 2: Definitions for qualitative ratings used in LLM storytelling evaluation

Rating	Definition
High	Model consistently performed very well with coherent, user-aligned, and appropriate responses across turns, without notable errors.
Moderate	Model performed reasonably well but had occasional minor issues such as slight incoherence, verbosity, or minor deviations from expected behavior.
Low	Model frequently exhibited problems such as irrelevance, inconsistency, disobedience to instructions, or unnatural phrasing that impacted the storytelling quality.

Human Study Design & Future Work

We aim to test two hypotheses:
- **H1:** Participants' mental health will improve over the course of the study.
- **H2:** LLM-integrated dynamic storytelling will yield greater psychological benefits than static storytelling.

We are conducting a within-subjects study to compare between two storytelling modes:
- **Dynamic Mode**: Misty uses an LLM for adaptive storytelling
- **Static Mode**: Misty selects from pre-written stories, unchanged by input

Participants interact with Misty in their dorm rooms over a one-week period. Each day, they complete mood surveys (PANAS), track their well-being (CSSWQ), and engage in at least three sessions. Psychological well-being (RPWS) is measured throughout the week. After the study, participants complete interviews to share their experiences and feedback.

Pre-Study
RPWS
CSWWQ

Daily Sessions
Daily sursey with
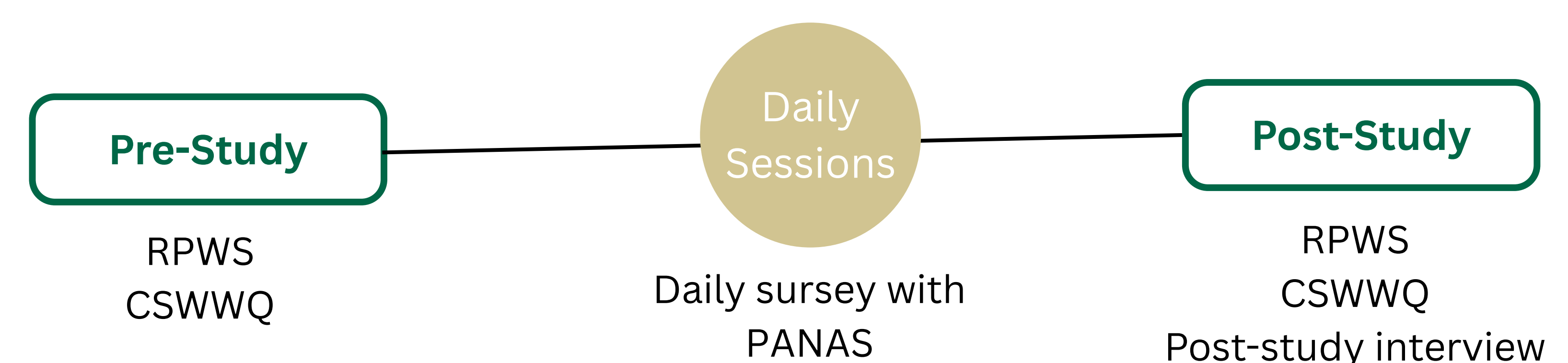PANAS

Post-Study
RPWS
CSWWQ
Post-study interview

Fig 3: Study timeline showing assessment points across the three phases: Pre-Study (baseline well-being surveys), Daily Sessions (mood tracking with PANAS), and Post-Study (final assessments and interviews).

Preliminary findings

Preliminary conversations suggest that Misty's dynamic storytelling fosters emotional openness and encourages self-reflection. Participants appreciated Misty's expressive behaviors and interactive responses. Next steps include analyzing long-term mental health outcomes and enhancing personalization in story generation.