GROUP 8

# AI Therapy Assistant

## Objective

*This study developed an AI therapy assistant to address academic stress among MSBA students, leveraging state-of-the-art NLP techniques to provide scalable, empathetic support while ensuring clinical safety.*

## Methodology

*Fine-tuned DistilBERT on a custom dataset (msba_ai_therapy_dataset) to classify emotions (e.g., stress, PTSD) and risk levels (high/low-risk).*
*o Compared zero-shot BART with fine-tuned DeBERTa, selecting the latter for superior emotion detection (F1: 0.74 vs. 0.56).*

## Core Techniques

### 1 Fine-Tuning & LoRA:

- Adapted DistilBERT using Low-Rank Adaptation (LoRA), freezing the base model while training only small, low-rank matrices (rank=16). This reduced GPU memory usage by 40% while maintaining 79% accuracy.
- Focused training on academic stress phrases (e.g., "extension request denied," "failed group project") from our custom dataset.

### 2 Prompt Engineering:

- **Emotion Specific Prompts:** Designed 12 emotion-specific prompt templates for FLAN-T5 (e.g., for "stress": *"Respond as a peer mentor: 1) Validate feelings 2) Suggest one 5-minute coping strategy"*).
- **Response Constraints:** Implemented response constraints: max 128 tokens, banned clinical terms ("diagnosis," "prescribe") to avoid overreach.

### 3 Risk Mitigation:

- Deployed regex patterns + classifier confidence scores (threshold: 0.85) to detect crises.
- Pre-loaded Warwick-specific resources: Wellbeing appointment links, 24/7 helpline numbers.

### 4 Human-in-the-Loop:

- Built review queue (JSON-based) prioritizing:
  a. High-risk predictions
  b. Low-confidence classifications (<0.6)
  c. Repeated user distress signals
- Included clinician override capability via simple REST API.

## Evaluations

- Achieved **79% accuracy** in risk classification and 74% F1-score in emotion detection.
- Confusion matrices revealed DeBERTa's strength in distinguishing nuanced emotions (e.g., sadness vs. loneliness).

## Key Innovations

- Hybrid architecture combining fine-tuned classifiers (emotion/risk) with generative FLAN-T5 for dynamic replies.
- ·Academic-specific prompt templates and safety protocols tailored to student stressors (e.g., assignment deadlines, visa anxiety).

## Further Developments

1. Adopt a multi-task transformer architecture with additional annotated data and GPU resources.
2. Expand dataset with non-English inputs and diverse student demographics.
3. Integrate multimodal input (e.g., voice tone analysis) for richer context.