School of Electronic Engineering and Computer Science
Queen Mary University of London

MSc PROJECT DEFINITION
2024-25

This project definition must be undertaken in consultation with your supervisor. The feasibility of the project should have been assessed and the project aims should be clearly defined.
Submission of this document implies that you have discussed the specification with your supervisor.

Project Title: EmoHeal: A Psychological Healing System Based on Fine-Tuned LLM and Multimodal Generation

Supervisor: Huan Zhang

Student name: Xinchen Wan

Student e-mail: ec24723@qmul.ac.uk

PROJECT AIMS:

State the design, development or research challenge (problem) that the project aims to solve.

Existing Challenges:

1. Current healing content generation systems (e.g., music playlists, motivational videos) rely on predefined templates and cannot dynamically adjust content based on users' real-time emotional states.
2. Existing tools generate text, audio, and images independently, resulting in a lack of **multimodal consistency** and fragmented therapeutic experiences.

Project Aims:
To design a **user-centered system** that dynamically generates emotion-adaptive healing content (music, voice, images) based on real-time input:

- **Core Focus**: Music generation as the primary therapeutic module, leveraging domain expertise in sound and music computing.
- **Integration Goal**: Seamlessly combine multimodal outputs into a cohesive video output.

PROJECT OBJECTIVES:

List a series of objectives you need to achieve to fulfil the aims of your project.

1. Emotion Analysis & Dynamic Adaptation

Purpose: Fine-tune open-source LLM models (e.g., Mistral-7B) to analyze user input and generate **hidden emotion tags** (e.g., "anxiety," "calmness") and encouraging text (e.g., "Take a deep breath").
Course Alignment:

- Apply NLP and LLM fine-tuning techniques from the Deep Learning for Audio and Music course.
- Optional integration of ontology reasoning from the Data Semantics course project.

---

2. Core Music Generation

Purpose: Generate emotion-adaptive therapeutic music based on LLM emotion tags using **text-to-music models** (e.g., MusicGen).
Course Alignment:

- Leverage music informatics principles (from Music Informatics course).

---

3. Multimodal Content Generation & Integration

Purpose:

- **Voice Synthesis**: Directly invoke pre-trained voice models from the Computational Creativity course to generate voice from hidden text.
- **Image Generation**: Create emotion-aligned images (e.g., "anxiety" → dark-toned abstract art) based on emotion tags.
- **Video Synthesis**: Use FFmpeg to synchronize audio (music + voice) and images into a 30-second video.

Course Alignment:

- Integrate computational creativity techniques (voice synthesis module from course projects).

---

4. System Development & Evaluation

Purpose:

- Develop a **Gradio-based interactive interface** (supporting text/voice input and video output).
- Evaluate system performance using **objective metrics** (e.g., alignment of music features with emotion tags) and **subjective user feedback** (e.g., Likert-scale surveys).

Course Alignment:

- Apply user interaction design principles from Computational Creativity.

State how your project will be aligned with the learning outcomes of your programme of study.

Alignment with Programme Learning Outcomes

- **Technical Competence**:
    - Advanced application of music generation, signal processing, and multimodal integration.
- **Research & Innovation**:
    - Novel integration of **LLM-based emotion analysis** with domain-specific music generation techniques.

METHODOLOGY:

Describe the various steps that you intend to follow for you to achieve your project aims.

To achieve the project objectives, the following steps will be executed:

---

1. Data Preparation & Preprocessing

- **Emotion-Music Dataset Construction**:
  - Curate music clips with emotion labels from public datasets (e.g., Free Music Archive) and supplement with GPT-4-generated descriptive text (e.g., "calm piano piece").
  - **Audio preprocessing**:
    - Normalize volume, segment into **30-second clips**, and extract features (e.g., tempo, tonality).
- **Voice Data Integration**:
  - Reuse pre-trained voice models from the Computational Creativity course without additional data collection.

---

2. Emotion Feedback LLM Fine-Tuning

- **Model Selection & Training**:
  - Use **Mistral-7B** with **LoRA** (Low-Rank Adaptation) for efficient fine-tuning.
  - **Input**: User text (e.g., "I feel anxious").
  - **Outputs**:
    - **Hidden encouraging text** (e.g., "Close your eyes and listen to the music") to drive downstream modules.
    - **Structured emotion labels** (e.g., "anxiety") and music feature tags (e.g., "slow_tempo").
- **Tools**: Hugging Face Transformers, PEFT (Parameter-Efficient Fine-Tuning) library.

---

3. Core Music Generation

- **Model**:
  - Use **Meta MusicGen** to generate music based on LLM emotion labels + **additional prompts** (e.g., "nature sounds + piano").

- **Output**: 30-second WAV audio file.

---

4. Multimodal Content Generation

- **Voice Synthesis**:
    - Directly invoke the course project's pre-trained voice models to generate voice from hidden text (MP3 format).
    - **Options**:
        - **Default voice**: Randomly selected pre-stored voices.
        - **Voice cloning**: Triggered when users upload reference audio (requires ethical review).
- **Image Generation**:
    - Use **Stable Diffusion XL** with emotion tags + keywords (e.g., "serene forest") to generate PNG images.

---

5. Video Synthesis & System Integration

- **Audio-Visual Synchronization**:
    - Use **FFmpeg** to combine music, voice, and images into a 30-second MP4 video, ensuring audio-visual synchronization.
- **Interactive Interface**:
    - Develop a **Gradio-based frontend**:
        - Supports text input, voice upload, video preview, and download.

---

6. Evaluation & Optimization

- **Objective Metrics**:
    - **Music feature alignment**: Match between generated music features (e.g., tempo) and emotion labels.
    - **Video generation latency**: Target ≤10 seconds.
- **Subjective Evaluation**:
    - Recruit users to assess therapeutic efficacy and multimodal coherence via **Likert-scale surveys** (1–5 points).

PROJECT MILESTONES

Indicate what measurable/tangible components you will produce as part of this project. This may take the form of deliverable document(s) or developmental milestones such as a working piece of software/hardware.

Milestone	Deliverables	Timeline
1. LLM Fine-Tuning & Emotion Analysis	- Fine-tuned Mistral-7B model (.bin) - Emotion tag mapping documentation	Weeks 1-2
2. Music Generation MVP	- MusicGen generation script (.py) - Post-processing rules (LibROSA)	Week 3
3. Multimodal Module Integration	- Voice synthesis interface (calls course project models) - Stable Diffusion image generation script	Week 4
4. Video Synthesis & Gradio Prototype	- FFmpeg automation script (.sh) - Interactive Gradio interface (input → video output)	Weeks 5-6
5. User Testing & Optimization	- User test report (PDF) - Code optimization (latency ≤15 seconds)	Weeks 7-9
6. Final Delivery	- Full code repository (GitHub) - Technical report (methodology + results) - Demo video (MP4)	Weeks 10-12

REQUIRED KNOWLEDGE/ SKILLS/TOOLS/RESOURCES:

Indicate as far as possible the skills that are required for you to undertake this project. Also include any software, hardware or other tools or resources that you believe you will need.

1. Knowledge & Skills

- **Core Technologies**:
  - NLP and LLM fine-tuning (Hugging Face ecosystem).
  - Music generation and signal processing (LibROSA, MusicGen).
  - Multimodal system integration (FFmpeg, Gradio).
- **Supporting Skills**:
  - Basic frontend development (Gradio interface design).
  - User research and data analysis (Likert-scale design).

---

2. Tools & Software

- **Development Tools**:
  - Python 3.10+, PyTorch, Hugging Face Transformers, Diffusers.
  - Audio processing: LibROSA, FFmpeg.
  - Image generation: Stable Diffusion XL, Diffusers.
- **Course Project Reuse**:
  - Pre-trained voice models from the Computational Creativity course (API or local access).

---

3. Hardware & Resources

- **Hardware**:
  - GPU: School-allocated A100 80G (for LLM fine-tuning and MusicGen inference).
  - Storage: 300GB SSD (for music datasets and generated content).
- **Datasets**:
  - Music: Free Music Archive (FMA) + GPT-4-generated descriptive text.
  - Voice: Pre-trained models and samples provided by the course project.

TIMEPLAN

This can be a GANTT chart submitted with this document or a list of tasks, milestones and deliverables with timings.

```
Gantt Chart (Weekly Breakdown)
Week | Task Progress
-----|-------------------------------------------------------------------------------
1    | [=====.................] Data preparation (25%)
2    | [==========............] LLM fine-tuning (50%)
3    | [===============.......] Music generation (75%)
4    | [====================...] Voice/image modules (90%)
5    | [========================] Video synthesis + Gradio prototype (100%)
6    | [========================] System integration testing (100%)
7    | [=====.................] User testing preparation (25%)
8    | [==========............] User preliminary testing (50%)
9    | [===============.......] User final testing (75%)
10   | [====================...] Ethical review + documentation (90%)
11   | [========================] Final optimization (100%)
12   | [========================] Project delivery (100%)
```

Key Deliverables

1. **Code & Models**:
    1. Fine-tuned LLM model (Mistral-7B + LoRA).
    2. MusicGen generation scripts.
    3. Gradio interactive interface (supports video preview/download).
2. **Documentation & Reports**:
    1. Technical report (methodology, test results, code documentation).
    2. User manual (Gradio usage guide).
    3. Ethical review documentation (if required).