**Project Title:**

**<titles>: Interactive Narrative-Driven Music Generation Using RAG and Hybrid AI Models**

**Project Deliverables**

**1. Problem Statement**

An AI-driven system that uses Retrieval-Augmented Generation (RAG) to generate interactive narratives and produces emotionally coherent music using a hybrid Transformer-Diffusion model.

**2. Abstract & Scope**

EmotionSync is an AI-powered system that generates dynamic, emotion-based music from interactive user prompts. It integrates RAG for real-time narrative generation and a hybrid music generation model combining Transformers for long-term coherence and Diffusion for structured composition. This system aims to enhance applications in gaming, therapy, storytelling, and personalized music experiences.

**Scope:**

* Generates interactive narratives from text-based user prompts.
* Maps the narrative’s emotional context to musical expressions.
* Uses pre-trained models for efficient training and fine-tuning.
* Supports integration into interactive media applications.

**3. Objectives**

1. **Develop an interactive AI system** that transforms textual prompts into structured narratives using RAG.
2. **Generate emotion-driven music** using a hybrid Transformer-Diffusion model.
3. **Ensure coherence** between narrative-driven emotion embeddings and musical outputs.
4. **Optimize performance** by using pre-trained models and fine-tuning on emotion-labeled music datasets.
5. **Evaluate system effectiveness** using both objective (audio similarity, structure) and subjective (listener surveys) metrics.

**4. Literature Survey**

| **Paper Title** | **Approach** | **Methodology** |
| --- | --- | --- |
| **A Study of Emotion-Based Music Generation Using Diffusion Model** | Uses a diffusion model for emotion-conditioned music generation. | Classifier-guided diffusion to control musical attributes aligned with specific emotions. |
| **MusicEmo: Transformer-Based Intelligent Approach Towards Music Emotion Generation and Recognition** | Uses a Transformer model for emotion recognition and generation. | Theme-based conditioning approach for emotional coherence. |
| **Semi-Supervised Emotion-Driven Music Generation Using Gaussian Mixture VAEs** | Uses VAEs to reduce reliance on labeled data while improving emotion control. | Disentangles rhythm and tonal features to allow controlled emotion-based generation. |
| **RAG for Contextual Music Generation** | Integrates RAG to retrieve relevant contextual data for music generation. | Uses knowledge retrieval to enhance the emotional coherence of AI-generated music. |
| **Contrastive Learning for Multi-Modal Music Understanding** | Applies contrastive learning for music-text associations. | CLIP-style learning to improve emotion-music alignment. |
| **Diffusion Models for Symbolic Music Generation** | Uses U-Net-based diffusion models for symbolic music. | Trains on symbolic representations like MIDI for structured outputs. |
| **Transformers for Long-Term Music Generation** | Applies attention-based mechanisms for structured composition. | Uses multi-head self-attention to capture dependencies in long musical sequences. |
| **Neural Music Composition via Pre-Trained Transformers** | Fine-tunes pre-trained music models for emotional expression. | Uses emotion embeddings as conditioning for music generation. |
| **Retrieval-Augmented Music Generation for Storytelling** | Uses retrieval-based augmentation for interactive music composition. | RAG pipeline retrieves thematic music references before generating new compositions. |

**5. Implementation Details**

**A. Data Collection & Preprocessing**

* **Datasets:** EMOPIA, MediaEval’s Emotify, VGMIDI (for emotion-labeled MIDI files).
* **Preprocessing:** Convert raw audio to MIDI and extract features (tempo, pitch, harmony) using **Librosa**.

**B. Narrative Generation with RAG**

* **Model:** Use **RAG (Facebook AI)** to generate structured narratives from user prompts.
* **Workflow:**
  1. User inputs a prompt or scenario.
  2. RAG retrieves relevant contextual information and generates an enriched storyline.
  3. The emotional state of the narrative is extracted (Valence-Arousal model).

**C. Music Generation using Hybrid Model**

* **Transformer Component:** Handles long-term coherence (e.g., GPT-2 or Music Transformer).
* **Diffusion Component:** Enhances structural stability and style (U-Net-based diffusion).
* **Workflow:**
  1. The emotional embeddings from RAG are fed into the hybrid model.
  2. The Transformer models generate initial MIDI sequences.
  3. The Diffusion model refines the generated sequences for better structure.

**D. Post-Processing & Evaluation**

* **Refinement:** Beat alignment, melody correction using MIDI post-processing tools.
* **Evaluation:**
  + **Objective:** Pitch stability, note density, rhythm accuracy.
  + **Subjective:** User surveys on emotional accuracy and musical coherence.

**6. High-Level Design**

**System Flow:**

1. **User Input** → Provides a narrative prompt.
2. **RAG Model** → Retrieves and generates a structured interactive narrative.
3. **Emotion Extraction** → Extracts emotion embeddings from the narrative.
4. **Hybrid Music Model** → Generates emotion-aligned music using Transformer + Diffusion.
5. **Post-Processing** → Melody correction, beat alignment.
6. **Output** → Plays or exports generated music.

**7. Low-Level Design**

**Module 1: Narrative Generation (RAG)**

* Uses pre-trained RAG to fetch and generate narrative data.
* Outputs structured story elements with emotional tags.

**Module 2: Emotion Embedding Extraction**

* Extracts features like valence-arousal mapping from the narrative.
* Uses sentiment analysis models (e.g., RoBERTa, BERT) to derive emotions.

**Module 3: Music Generation**

* Uses a Transformer to generate the MIDI skeleton.
* Diffusion model refines structure and composition.

**Module 4: Post-Processing**

* Adjusts tempo, rhythm, and ensures coherence in MIDI sequences.
* Converts MIDI into high-quality audio using a synthesizer (FluidSynth, MuseScore).

**8. Deliverables**

| **Phase** | **Deliverables** |
| --- | --- |
| **Phase 1: Research & Planning** | Literature survey, dataset selection, high-level design. |
| **Phase 2: Model Implementation** | Implement RAG for narrative generation, Transformer-Diffusion model for music. |
| **Phase 3: Integration & Testing** | Connect RAG with emotion extraction and music generation pipeline. |
| **Phase 4: Evaluation & Optimization** | Objective & subjective evaluations, model fine-tuning. |
| **Phase 5: Final Deployment** | Deploy interactive system with UI/API for real-time input. |

**9. Complexity & Feasibility**

| **Aspect** | **Complexity Level** | **Challenges** | **Simplification Strategy** |
| --- | --- | --- | --- |
| **Training RAG + Music Model** | 🔥 High | Requires multi-modal dataset & alignment | Use pre-trained RAG, fine-tune for domain-specific narratives |
| **Ensuring Emotion-Music Coherence** | 🚀 Very High | Mapping textual emotion to music is non-trivial | Use Valence-Arousal emotion embedding models |
| **Computational Efficiency** | ⚡ Medium | Hybrid model inference can be slow | Optimize inference with model distillation |

**Feasibility:**

* Using **pre-trained models** for RAG and Music Transformer makes this project feasible.
* Training from scratch would be impractical for small-scale execution.
* The major challenge is aligning **narrative structure → emotion embedding → musical generation**, but this can be handled through **transfer learning** and **emotion-conditioned models**.

**Conclusion**

This project integrates **RAG for contextual storytelling** and **a hybrid AI model (Transformer + Diffusion) for music generation**, allowing **emotionally-coherent** compositions from user-provided prompts. With optimization strategies such as **pre-trained models, fine-tuning, and efficient embeddings**, the system is **scalable** and **feasible** while maintaining high-quality, expressive outputs.

**Problem Statement:**  
Current systems for interactive narrative and music generation lack seamless integration, resulting in emotionally disjointed experiences. The absence of real-time alignment between dynamically generated narratives and their musical counterparts undermines immersion in applications like gaming and therapy. A critical challenge lies in mapping textual emotions to structured musical features efficiently. This gap demands an AI system integrating RAG for narrative coherence and a hybrid Transformer-Diffusion model for emotion-driven music composition.

give the objectives of the project

Thought for 11 seconds

Alright, let me break down the user's request. They want the objectives of the project after I provided a problem statement. The user probably needs clear, concise goals that address the issues outlined.

First, I need to recall the project details. The main aim is integrating RAG for narratives and a hybrid AI model for music. The problem statement highlighted coherence issues between narratives and music, real-time alignment, and emotion mapping.

So, the objectives should tackle each part of the problem. The first objective would be developing the RAG system for narratives. Then, creating the music model. Ensuring they work together is crucial, so coherence between emotion and music is another objective. Optimization and evaluation are also key for practicality and validation.

I should structure each objective to directly respond to the problem points. Using action verbs like "Develop," "Design," "Ensure," "Optimize," and "Evaluate" makes them clear and actionable. Need to keep each objective specific and aligned with the project's scope. Let me check if each objective addresses a part of the problem without overlapping. Also, make sure they're concise and numbered for clarity.

**Project Objectives:**

1. **Develop a RAG-driven system** to generate interactive, contextually rich narratives from user prompts in real-time.
2. **Design a hybrid Transformer-Diffusion model** to produce structured, emotion-coherent music aligned with narrative-driven emotional embeddings.
3. **Ensure seamless alignment** between textual emotion extraction (e.g., valence-arousal) and musical features (tempo, harmony, rhythm).
4. **Optimize computational efficiency** using pre-trained models and fine-tuning strategies for scalable deployment.
5. **Evaluate system performance** through objective metrics (audio structure) and subjective user feedback (emotional resonance).

An imersed output with music and story scenario based on user input make this as 1 of the objectives

**Project Objectives:**

1. **Develop a RAG-driven system** to generate interactive, contextually rich narratives from user prompts in real-time.
2. **Design a hybrid Transformer-Diffusion model** to produce structured, emotion-coherent music aligned with narrative-driven emotional embeddings.
3. **Deliver an immersive output** that seamlessly integrates dynamically generated music and story scenarios based on user input.
4. **Ensure seamless alignment** between textual emotion extraction (e.g., valence-arousal) and musical features (tempo, harmony, rhythm).
5. **Optimize computational efficiency** using pre-trained models and fine-tuning strategies for scalable deployment.
6. **Evaluate system performance** through objective metrics (audio structure) and subjective user feedback (emotional resonance).

give the proposed methodology for the project

**Proposed Methodology**

The project will follow a structured, multi-phase methodology to achieve its objectives. Below is a detailed breakdown of the proposed approach:

**1. Data Collection and Preprocessing**

* **Datasets:**
  + **Narrative Data:** Use publicly available story datasets (e.g., ROCStories, Aesop’s Fables) for training RAG.
  + **Music Data:** Leverage emotion-labeled datasets like EMOPIA, VGMIDI, and MediaEval’s Emotify for training the hybrid music generation model.
* **Preprocessing:**
  + Convert raw audio to MIDI format using tools like Librosa.
  + Extract musical features (tempo, pitch, harmony) and emotion labels (valence-arousal) for training.
  + Tokenize narrative data for RAG and align it with emotional context.

**2. Narrative Generation with RAG**

* **Model:** Use a pre-trained RAG model (e.g., Facebook AI’s RAG) for narrative generation.
* **Workflow:**
  1. **User Input:** Accept a textual prompt (e.g., "a hero’s journey through a dark forest").
  2. **Context Retrieval:** RAG retrieves relevant contextual information from a knowledge base.
  3. **Narrative Generation:** Generate a structured, emotionally rich narrative enriched with retrieved context.
  4. **Emotion Extraction:** Use sentiment analysis models (e.g., RoBERTa, BERT) to extract emotion embeddings (valence-arousal) from the narrative.

**3. Music Generation with Hybrid Transformer-Diffusion Model**

* **Transformer Component:**
  + Use a pre-trained Transformer (e.g., Music Transformer, GPT-2) to generate MIDI sequences based on emotion embeddings.
  + Focus on long-term coherence and thematic consistency.
* **Diffusion Component:**
  + Use a U-Net-based diffusion model to refine the MIDI sequences for better structure and style.
  + Ensure alignment with the emotional tone of the narrative.
* **Workflow:**
  1. Feed emotion embeddings from RAG into the Transformer to generate a MIDI skeleton.
  2. Refine the MIDI sequences using the Diffusion model for enhanced structure and emotional expressiveness.

**4. Post-Processing and Integration**

* **MIDI Refinement:**
  + Use tools like MuseScore or MIDI ToolKit for beat alignment, melody correction, and tempo adjustment.
* **Audio Synthesis:**
  + Convert refined MIDI sequences into high-quality audio using synthesizers like FluidSynth or Ableton Live.
* **Integration:**
  + Combine the generated narrative and music into a cohesive output (e.g., a playable audio file with a synchronized story display).

**5. Evaluation and Optimization**

* **Objective Evaluation:**
  + Measure pitch stability, note density, rhythm accuracy, and structural coherence of the generated music.
  + Use metrics like Fréchet Audio Distance (FAD) to compare generated music with ground truth.
* **Subjective Evaluation:**
  + Conduct user surveys to assess emotional accuracy, narrative-music coherence, and overall immersion.
* **Optimization:**
  + Fine-tune models using feedback from evaluations.
  + Apply techniques like model distillation to improve computational efficiency.

**6. Deployment and User Interaction**

* **User Interface (UI):**
  + Develop a simple, intuitive UI for users to input prompts and receive generated narratives and music.
* **API Integration:**
  + Provide an API for integration into applications like gaming engines, virtual reality platforms, or therapeutic tools.
* **Real-Time Interaction:**
  + Ensure low-latency generation for real-time user interaction.

**Summary of Methodology**

1. **Data Preparation:** Collect and preprocess narrative and music datasets.
2. **Narrative Generation:** Use RAG to generate emotionally rich narratives from user prompts.
3. **Music Generation:** Employ a hybrid Transformer-Diffusion model to create emotion-coherent music.
4. **Post-Processing:** Refine and synthesize the output for high-quality audio and narrative integration.
5. **Evaluation:** Assess system performance using objective and subjective metrics.
6. **Deployment:** Develop a user-friendly interface and API for real-time interaction.

This methodology ensures a systematic approach to achieving the project’s objectives while addressing technical and computational challenges.