From Recognition to Expression:

A Multimodal Emotion-Aware AI with Personalized Voice Cloning

Introduction - Recent times have made us more dependent on technology than ever, and AI has quietly become part of our everyday routines — from recommending songs to answering questions late at night. Yet, as shown in a recent study by the **University of Bamberg, Germany, titled "Fearful Falcons and Angry Llamas: Emotion Category Annotations of Arguments by Humans and LLMs"**, today's AI systems still struggle with emotional nuance. The paper reveals that Large Language Models (LLMs) often misinterpret or entirely overlook emotional intent in argumentative texts — a gap that becomes especially visible when comparing their annotations to those of humans. This highlights a critical shortcoming in how AI processes the 'how' of communication, not just the 'what'.

For humans, emoting is like breathing — we do it naturally, without much thought. But many users forget, or don't even realize, that the AI they're speaking to is just a machine. This trend is particularly concerning in the context of rising social isolation, where individuals increasingly form emotional attachments to AI systems, often attributing human-like qualities to technologies that are not equipped to reciprocate or understand emotional depth. Without transparency from tech providers, people may begin to rely on these systems for comfort or connection, unintentionally blurring the lines between emotional support and artificial interaction.

That's where Emotion-LLaMA steps in. While not a perfect fix, it's a move toward AI that feels a little more human. By blending facial micro-expressions, voice tone, and text context, Emotion-LLaMA learns to perceive and reason about emotions, rather than simply labeling them.

Whether it's aiding recovering patients who need around-the-clock support or making virtual assistants more emotionally intelligent, Emotion-LLaMA pushes AI closer to truly understanding us — not just as users, but as humans.

Goal

To develop a multimodal emotion recognition and reasoning system that accurately identifies and explains emotions across various contexts by combining audio, video, and textual data using a tuned LLM framework.

Objectives

1. **Develop the Emotion-LLaMA model for audio, video, and text processing** - Emotion-LLaMA integrates audio, visual, and textual inputs using specialized encoders. These features are projected into a shared embedding space and tokenized for processing by a unified language model, enabling deep emotional understanding from multimodal content.
2. **Apply instruction tuning to align multimodal features with emotion reasoning tasks** -
   The model is fine-tuned using instruction-based prompts that guide it to not only recognize emotions but explain them using evidence from different modalities. Training includes both coarse- and fine-grained data from the MERR dataset and others, allowing the model to handle complex reasoning tasks and follow natural-language instructions effectively.
3. **Benchmark the model against SOTA models on multiple datasets and tasks** – Evaluate on various datasets, using metrics such as F1 score, UAR, WAR, and reasoning overlap scores.
4. **Provide emotion reasoning, not just classification, for enhanced interpretability** – Beyond simply labeling emotions, Emotion-LLaMA generates detailed reasoning by combining vocal cues, facial expressions, textual dialogue, and contextual background.
5. **Build an Emotional Text-to-Speech (TTS) System with Voice Cloning** – (Ambitious objective)
   Develop a TTS module that takes a text input and a target emotion and synthesizes emotional speech in the user's own voice using voice cloning techniques. This system extends the utility of Emotion-LLaMA by enabling expressive, personalized speech generation, creating a complete pipeline from emotion recognition to emotional expression.

Emotion-LLaMA Output	Used in TTS as
• Emotion label (e.g., "anger") • Original or generated text • Optional: emotional reasoning or modified phrasing	• Emotion control signal or embedding • Spoken content • Prosody control or expressive text prompt

Knowledge and Skills Requirement

- Machine Learning / Deep Learning, Natural Language Processing (NLP), Computer Vision

- Speech Processing (ASR & SER)
- Multimodal Fusion Techniques, Transformer Architectures / LLMs
- Python, PyTorch, HuggingFace, OpenCV

Software & Tools

- **Frameworks**: PyTorch, HuggingFace Transformers, OpenMMLab
- **Audio Models**: HuBERT / WavLM (will be more robust to noise), Qwen-Audio
- **Vision Models**: MAE, VideoMAE, EVA
- **Language Model**: LLaMA-3 / LLaMA-2 with LoRA
- **Toolkits**: OpenFace (for AU extraction), SentencePiece (tokenization)
- **Demo & Hosting**: HuggingFace Spaces, GitHub
- Text-to-Speech & Voice Cloning - Coqui TTS
- Explore the integration of AudioFlamingo as a future use case.

Datasets –

MERR - Google Drive

Emotion-LLaMA/MERR/README.md at main · ZebangCheng/Emotion-LLaMA

Dataset	Source	Characteristics	Justification
MERR	Created by authors from MER2023	28,618 coarse & 4,487 fine-grained samples, multimodal annotations (audio, video, subtitles, emotion descriptions)	Rich, diverse emotional contexts for training Emotion-LLaMA
MER2023	MER Challenge	Multi-label, semi-supervised, AVT modalities	Benchmark for emotional recognition tasks
MER2024	MER Challenge	Noisy and open-vocabulary scenarios	Evaluates model's robustness and open-set recognition
DFEW	Academic Dataset	Emotion in wild movie clips	Tests generalization and real-world performance
EMER	Emotion reasoning benchmark	Includes emotion triggers and explanation prompts	Measures reasoning ability across modalities

Research Papers (Key References)

- HuBERT: "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units" by Wei-Ning Hsu et al.
- WavLM: "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing" by Shuai Wang et al.
- Qwen-Audio: "Qwen-Audio: Advancing Universal Audio Understanding via Audio-Language Pretraining" by Yifan Xu et al. MAE / VideoMAE: *Masked Autoencoders for Visual Understanding*
- MAE: "Masked Autoencoders Are Scalable Vision Learners" by Kaiming He et al.
- VideoMAE: "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training" by Zhan Tong et al.
- EVA: "EVA: Exploring the Limits of Masked Visual Representation at Scale" by Yuxin Wang et al.
- LLaMA-3: "The Llama 3 Herd of Models" by Aaron Grattafiori et al.
- LoRA: "LoRA: Low-Rank Adaptation of Large Language Models" by Edward J. Hu et al.
- OpenFace: "A Non-Invasive Approach for Facial Action Unit Extraction and Its Application in Pain Detection" by Mondher Bouazizi et al.
- SentencePiece: "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing" by Taku Kudo and John Richardson.
- **AudioFlamingo**: "Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Reasoning Capabilities" by Luyu Wang et al.
- LLaVA / MiniGPT-v2: *Visual Instruction Tuning*
- GPT-4V: *Vision-language alignment benchmarks*
- AffectGPT: *Emotion Reasoning with LLMs*
- MER2023/2024 Challenge Papers
- Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning
- EMER Dataset Paper: *Explainable Multimodal Emotion Reasoning*

Plan of Action

Phase 1: Background Research - Study existing MLLMs and their limitations for emotion recognition. Review HuBERT, MAE, and LLaMA instruction tuning papers.

Phase 2: Dataset Preparation - Analyze and experiment with the MERR dataset. Understand annotation pipeline: AU detection (OpenFace), audio descriptors, multimodal description generation.

Phase 3: Model Development - Implement Emotion-LLaMA with separate encoders:

HuBERT for audio; MAE, VideoMAE, EVA for vision; LLaMA for language + reasoning - Apply linear projection to align modalities.

Phase 4: Training - Pretraining: On coarse-grained MERR samples. Instruction Tuning: On fine-grained MERR + MER/DFEW datasets.

Phase 5: Evaluation - Run experiments on:

MER2023 (F1 score)

MER2024 (Noise, OV)

DFEW (UAR, WAR)

EMER (Clue & Label Overlap)

Compare against other MLLMs (Video-LLaMA, GPT-4V, etc.)

Phase 6: Analysis & Visualization

Conduct ablation studies for each encoder. Perform qualitative reasoning analysis. Visualize output distributions, error cases.