Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.085437

ARTICLE

A Comparative Study of Audio-Language Models for Speech Emotion Recognition in Spanish

Jorge Gómez-Navalón, Ronghao Pan, Tomas Bernal-Beltrán, José Antonio García-Díaz* and Rafael Valencia-García

Informatics and Systems, Universidad de Murcia, Murcia, Spain

*Corresponding Author: José Antonio García-Díaz. Email: joseantonio.garcia8@um.es

Received: 11 May 2026; Accepted: 02 July 2026; Published: 27 July 2026

ABSTRACT: Traditionally, speech emotion recognition has relied on supervised models that require task-specific training and annotated data. However, the recent emergence of audio-language models introduces a more flexible paradigm that enables multimodal reasoning through speech and natural language interaction. Nevertheless, their effectiveness for emotion recognition remains unclear. In this study, we evaluate audio-language models for speech emotion classification using the Spanish MEACorpus dataset and compare three approaches: prompt-based inference, embedding-based classification with lightweight classifiers, and instruction-tuned models with parameter-efficient fine-tuning plus a hybrid architecture based on class-specific confidence-driven routing. Our results show that the hybrid approach achieves the highest overall performance, reaching an 83.55% macro F1-score and an 84.37% weighted F1-score. Instruction tuning remains highly competitive, obtaining an 83.32% macro F1-score, which confirms the importance of supervised task adaptation for aligning ALMs with speech emotion recognition. We also include Whisper as a pretrained acoustic baseline to contextualize ALM-based representations against strong speech foundation models. Furthermore, the hybrid approach outperforms standalone embedding-based classification across all evaluated models, showing that class-specific confidence-driven routing can improve the use of embedding-based predictions. Although the best hybrid ALM configuration achieves competitive performance, it remains below the specialized MEACorpus baseline of 87.74% macro F1-score, indicating that general-purpose ALMs do not yet surpass highly specialized acoustic models for Spanish speech emotion recognition.

KEYWORDS: Speech emotion recognition; audio-language models; multimodal learning; instruction tuning; prompting; embedding-based classification

1 Introduction

Human emotions play a central role in communication, influencing how information is expressed and interpreted. The automatic recognition of emotions from speech, known as Speech Emotion Recognition (SER), has attracted increasing attention in affective computing [1]. SER has been applied in various fields, including human-computer interaction [2], virtual assistants, call center analytics and mental health monitoring systems [3,4], where an understanding of emotional states can inform decision-making processes.

In recent years, advances in deep learning and self-supervised learning have significantly improved SER performance, particularly through the use of transformer-based architectures and pretrained speech representations such as wav2vec 2.0 and HuBERT [5,6]. These approaches enable end-to-end learning from raw audio and reduce the need for handcrafted features. However, they typically require labeled

data and task-specific training, which limits their adaptability in low-resource scenarios and restricts their generalization capabilities. Recently, the emergence of Audio-Language Models (ALMs) has introduced a new paradigm for processing and reasoning over speech signals. By extending the capabilities of Large Language Models (LLMs) to the audio domain, ALMs integrate acoustic and linguistic information within a unified framework [7,8]. This enables more flexible interaction through natural language instructions and new inference strategies beyond traditional supervised classification.

In the context of SER, ALMs support three complementary approaches: (i) prompt-based inference, where models generate predictions in zero-shot or few-shot settings; (ii) embedding-based pipelines, where pretrained models serve as feature extractors combined with lightweight classifiers; and (iii) instruction tuning, where models are adapted to the task using parameter-efficient fine-tuning techniques [9]. These paradigms offer different trade-offs in terms of flexibility, efficiency, and performance. However, their comparative effectiveness in realistic settings remains unclear. This lack of systematic evaluation is particularly relevant given the inherent trade-offs of each approach. Prompt-based methods are flexible, yet they are often unstable and computationally expensive. Embedding-based approaches are efficient and scalable but may lack the expressive capacity of generative models. Instruction tuning improves task alignment but increases training complexity and resource requirements.

Despite recent advances and surveys on multimodal learning and ALMs [10,11], there is still no clear consensus on which paradigm is most suitable for SER under practical constraints. Specifically, existing studies lack a unified experimental comparison of these approaches, which makes it difficult to evaluate their respective strengths, limitations, and trade-offs in realistic scenarios.

This motivates the need for a systematic comparison of ALM-based strategies for SER rather than assuming that larger or more flexible multimodal models will automatically perform well on emotion recognition tasks. SER requires sensitivity to fine-grained acoustic and paralinguistic cues, while ALMs are often designed for broader audio-language understanding and instruction-following scenarios. Therefore, it is important to analyze how different adaptation paradigms behave under the same experimental conditions, especially in Spanish, where SER resources and evaluations remain more limited than in English. Such an analysis can clarify whether prompting, embedding-based classification, instruction tuning, or hybrid routing provides the most effective trade-off between flexibility, computational cost, and recognition performance.

In this work, we present a systematic empirical evaluation of ALMs for SER using the Spanish MEACorpus dataset [12]. We examine three complementary paradigms: (1) prompt-based inference under zero-shot and few-shot configurations, (2) embedding-based approaches with frozen ALMs and lightweight classifiers, and (3) instruction-tuned models adapted through parameter-efficient fine-tuning. Building on these paradigms, we propose a hybrid architecture based on confidence-driven routing. First, the approach performs prediction using embedding-based classifiers derived from different ALMs. Then, predictions with low confidence are selectively reprocessed using the best-performing instruction-tuned model, Flamingo, which acts as a robust fallback mechanism. This strategy combines the efficiency of embedding-based classification with the stronger task alignment of instruction-tuned inference. In addition, we include Whisper as a non-instructional pretrained acoustic baseline, using its encoder representations with the same lightweight classifier employed in the embedding-based experiments. Our study includes multiple state-of-the-art ALMs, such as Flamingo and Gemma, evaluated within a unified experimental framework.

The rest of the manuscript is organized as follows. Section 2 reviews the state of the art in SER, multimodal approaches, and ALMs. Section 3 describes the proposed system architecture. Section 4 details the experimental setup. Section 5 presents the results and discussion. Finally, Section 6 concludes the paper and outlines future research directions.

2 State of the Art

This section is organized into two subsections: the first focuses on SER (see Section 2.1) and the second to ALMs and prompting (see Section 2.2).

2.1 Speech Emotion Recognition

Emotion analysis is a broad research area within affective computing that aims to automatically identify and interpret human emotional states from various modalities, including text, speech, facial expressions, and physiological signals. Several psychological models have been proposed to characterize emotions. Two of the most widely adopted frameworks in computational studies are the discrete emotion taxonomy introduced by Ekman [13] and Plutchik's Wheel of Emotions [14]. Within this field, SER focuses specifically on detecting the emotional state of a speaker from speech signals by exploiting both acoustic and prosodic cues. SER has attracted considerable attention due to its applications in human-computer interaction [15], mental health monitoring [3,4], and affective computing systems [16,17].

Early approaches to SER relied on traditional machine learning techniques such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM), combined with handcrafted acoustic features including Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy [2,18]. These handcrafted descriptors remain important in classical SER pipelines, since emotional states are reflected through prosodic, spectral, and temporal variations in speech signals. Common features include MFCCs, pitch or fundamental frequency, energy, zero-crossing rate (ZCR), and other time- and frequency-domain descriptors, which capture complementary aspects of vocal effort, rhythm, and spectral structure [19,20]. Feature selection is also relevant in these pipelines, since redundant or irrelevant acoustic descriptors can increase computational complexity and degrade classification performance. In this line, threshold-based feature selection has been explored to obtain compact and discriminative feature subsets for multilingual or regional-language SER scenarios [20].

Despite the usefulness of handcrafted descriptors and feature selection, traditional SER pipelines still depend heavily on manual feature engineering and pipeline-based processing, which limits their ability to capture complex and high-level emotional patterns in speech [21]. To overcome these limitations, deep learning models have become the dominant paradigm in SER. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, enable the automatic extraction of hierarchical and temporal representations directly from speech signals [22,23]. These architectures significantly improve performance by learning feature representations directly from data rather than relying on handcrafted descriptors.

Building on these advances, transformer-based architectures have recently gained prominence due to their ability to model long-range dependencies and capture global contextual information in speech signals. Unlike RNNs, transformers rely on self-attention mechanisms, which allow for more effective modeling of temporal relationships and have demonstrated strong performance across a wide range of speech processing tasks [24]. Recent specialized SER architectures have also explored combinations of temporal modeling, attention mechanisms, and hierarchical feature extraction. For example, Zhang et al. proposed STACN, a sparse temporal-aware capsule network that combines multi-head attention, temporal-aware blocks, residual connections, and capsule modules to improve robustness, generalization, and reproducibility across multiple SER benchmarks [25]. However, these improvements do not fully address all limitations of SER, as deep learning-based SER models still face significant challenges, particularly in terms of generalization across datasets and speakers. Variability in recording conditions, speaker characteristics, and annotation inconsistencies often leads to degraded performance in cross-corpus scenarios. Furthermore, the reliance on labeled data can result in overfitting, particularly in low-resource settings [22].

Self-supervised pretraining has emerged as a major paradigm shift that addresses the limitations in speech representation learning. Models such as *wav2vec 2.0* [5], *HuBERT* [6], and *WavLM* [26] use large-scale unlabeled audio data to learn general-purpose representations that can then be fine-tuned for downstream tasks such as SER. In this direction, Chen et al. introduced MTLSER, a multi-task learning framework that uses wav2vec 2.0 as a shared acoustic representation layer and jointly optimizes speech emotion recognition with auxiliary automatic speech recognition and speaker identification tasks [27]. This illustrates the continued importance of pretrained acoustic models combined with task-specific training objectives for SER. This approach significantly reduces the dependence on annotated data while improving the robustness and transferability of the learned features. When combined with lightweight classifiers or traditional machine learning methods, these pretrained representations serve as fixed feature extractors. They demonstrate strong performance compared to handcrafted features [28,29]. However, these approaches remain primarily focused on acoustic modeling and may still struggle to capture higher-level contextual and semantic aspects of emotional expression. Recently, this paradigm has evolved to include the development of foundation models for speech, such as Whisper [30], which was trained at large scale for robust speech recognition and has also been explored as a source of transferable acoustic representations. These large-scale pretrained models are trained on diverse and heterogeneous data and can be adapted to multiple downstream tasks. This enables more flexible transfer learning strategies [26,30]. While promising, their application to SER remains an active area of research, particularly in scenarios where both acoustic and semantic understanding are required [11].

Overall, current SER approaches continue to be challenged by variability in speakers and recording conditions, as well as the inherent ambiguity of emotional expression [22,23]. Despite significant advances, purely acoustic models are still limited in their ability to capture the full complexity of emotion. The performance of SER systems is strongly influenced by the characteristics of available datasets, which vary in terms of language, recording conditions, and annotation protocols [31]. Widely used benchmarks such as EmoDB [32], IEMOCAP [33], and RAVDESS [34] are based on acted or semi-controlled recordings, which may limit their generalization to real-world scenarios. More recent datasets, such as MELD [35], incorporate more natural and spontaneous emotional expressions.

Despite the abundance of English datasets, resources for SER in Spanish remain limited. Existing datasets either include Spanish within multilingual collections [36] or are limited in size and recording conditions, such as EmoMatchSpanishDB [37]. In contrast, the Spanish MEACorpus [12] provides a multimodal dataset collected from real-world environments, making it a more suitable benchmark for evaluating SER systems under realistic conditions. However, even with improved datasets, purely acoustic approaches are limited in their ability to capture the full complexity of emotional expression.

Despite the progress achieved by acoustic-based approaches, SER alone remains limited in its ability to capture the contextual and semantic aspects of human communication [38]. In real-world scenarios, emotional expression is conveyed through multiple complementary modalities, including speech, text, facial expressions, and contextual cues. Therefore, multimodal emotion recognition naturally overcomes the inherent limitations of unimodal systems [39].

Multimodal integration, or the effective combination of signals from different modalities, remains a challenging problem. Temporal misalignment, heterogeneous feature spaces, and inconsistent semantic representations complicate the fusion process. These challenges are particularly pronounced in realistic conditions, where data may be noisy or incomplete [40,41]. Existing approaches typically rely on either early fusion, which combines features into a shared representation, or late fusion, which integrates modality-specific outputs [42,43]. More recent methods use attention mechanisms and multimodal transformers to model complex cross-modal interactions [44], although these approaches often require large amounts of well-aligned data and entail higher computational costs [45–47]. Overall, while multimodal approaches

provide richer representations of emotional expression, their effectiveness is limited by the difficulty of reliably integrating heterogeneous and incomplete data sources [48–50]. These limitations have motivated the exploration of more flexible paradigms that can jointly integrate perception and reasoning. One such paradigm is ALMs, which aim to unify acoustic and linguistic understanding within a single framework.

2.2 Audio-Language Models and Prompting

Recent advances in LLMs and multimodal learning have led to the emergence of ALMs, a class of multimodal systems that integrate speech encoders with language models to process and reason over speech signals [11,51]. Table 1 summarizes several ALMs and their main characteristics, highlighting differences in backbone architectures, audio encoders, and supported capabilities. All of these models support audio and textual inputs.

Table 1: Representative ALMs and their main characteristics.

Model	Backbone	Audio Encoder	Capabilities	Size	Context
SpeechGPT [52]	LLaMA	HuBERT	Generation, understanding	7B	2K
AudioGPT [53]	LLM-based	Multiple	Multi-task audio reasoning	–	–
Qwen2-Audio [54]	Qwen2	WavLM-like	Instruction-following	7B	32K
Gemma-4 [55]	Gemma	Proprietary	Efficient inference	4B	32K
Voxtral [56]	Custom LLM	Proprietary	Large-scale reasoning	24B	32K
MOSS-Audio [57]	MOSS	Custom	Instruction tuning	4B	8K

Unlike traditional SER approaches based on supervised classification, ALMs reformulate SER as an instruction-following task in which predictions are generated through natural language interaction, allowing flexible inference through prompting [53,58]. This paradigm is particularly relevant for SER because emotional understanding depends on acoustic cues and semantic content.

Typically, ALMs combine a pretrained audio encoder, such as wav2vec 2.0 [5], HuBERT [6], or WavLM [26], with a large language model backbone. Audio representations are projected into a shared representation space that is compatible with the language model, enabling joint processing of acoustic and linguistic information [52]. To align the modalities, parameter-efficient fine-tuning techniques such as LoRA are commonly used [59]. These techniques freeze the pretrained LLM weights and learn only a small set of low-rank adaptation parameters, substantially reducing computational cost.

Self-supervised audio models provide strong acoustic representations but are usually limited to feature extraction followed by task-specific classifiers. These classifiers often require large amounts of annotated data for downstream training, which makes the adaptation process costly and time-consuming [28]. In contrast, ALMs introduce reasoning capabilities through natural language interaction. However, recent studies suggest that multimodal LLMs may struggle to capture fine-grained acoustic and paralinguistic cues, such as prosody and intonation.

Prompting plays a central role in ALM-based SER, as it determines how the model interprets and reasons about speech inputs. In zero-shot prompting, the model relies solely on task instructions. In contrast, few-shot prompting incorporates labeled examples to improve task adaptation [60]. For this reason, in SER, prompt design is particularly critical due to the ambiguity of emotional expression, which can lead to variability in model predictions [61]. Techniques such as chain-of-thought prompting aim to improve interpretability by encouraging step-by-step reasoning. However, these techniques remain limited in their ability to capture fine-grained emotional cues, as demonstrated in studies such as [62,63].

Despite their flexibility, ALMs have several limitations. They may exhibit biases toward textual inputs, show high sensitivity to prompt design, and produce inconsistent or hallucinated outputs [64,65]. These challenges highlight the need for a systematic evaluation of prompting strategies and modality integration in SER tasks, particularly in the Spanish language [66].

3 System Architecture

In this work, we explore different architectural strategies for SER using ALMs. The proposed system processes multimodal inputs, combining audio signals and optional textual transcriptions. It also evaluates the impact of different modeling paradigms on performance.

Our approach is structured around three primary paradigms and an additional hybrid strategy that combines them. First, we examine prompt-based inference, in which pretrained ALMs are used to directly generate emotion labels from input prompts under zero-shot and few-shot configurations. This allows us to evaluate their ability to generalize without task-specific training. Second, we explore an instruction tuning strategy, adapting the same models using supervised conversational data. We reformulate the task as a dialogue between a user and an assistant to better align it with the classification objective. Third, we consider an embedding-based architecture, in which ALMs act as feature extractors producing dense representations of the input. These representations are then processed by a lightweight classification head based on a Multi-Layer Perceptron (MLP).

We introduce a hybrid architecture integrating embedding-based classification and instruction-tuned generative inference via a confidence-driven routing mechanism. In this framework, the embedding-based classifier associates predictions with confidence scores. High-confidence predictions are accepted directly according to the threshold associated with the predicted emotion class, whereas low-confidence cases are routed to the instruction-tuned Flamingo model for further inference. This design balances computational efficiency and predictive robustness by applying more complex inference only when necessary. Overall, the system provides a unified framework for systematically comparing prompting, fine-tuning, and representation-based strategies for emotion classification in multimodal audio scenarios. An overview of the proposed system is presented in Fig. 1, illustrating the different processing pipelines and their interactions.

3.1 Input Representation

The proposed system operates on multimodal inputs, which are composed of audio signals and, optionally, their corresponding textual transcriptions. Each input sample is associated with an audio segment and a transcription that is either manually or automatically generated. This allows us to explore both unimodal and multimodal configurations. Therefore, we define two primary input modes. In the audio-only setting, the model receives only the raw audio signal. In the audio + transcript setting, the input is enriched with the corresponding transcription, which is incorporated into the prompt alongside the audio. This multimodal formulation enables the model to leverage acoustic and linguistic cues for emotion recognition.

To ensure compatibility across different ALMs, inputs are formatted using a conversational structure. Specifically, each sample is represented as a sequence of messages. The first message is a system prompt that defines the task. The second message is a user prompt that includes the input data. Depending on the selected mode, the user prompt may contain either the audio signal alone or a combination of a textual transcription and the audio signal. In the audio+transcript setting, for example, the user prompt may include a textual instruction followed by the transcription and the associated audio segment. This enables the model to process both modalities jointly. This unified input representation allows us to evaluate different modeling strategies, such as prompt-based inference, fine-tuning, and embedding extraction, while maintaining a consistent interface. The proposed conversational format is shown in Table 2.

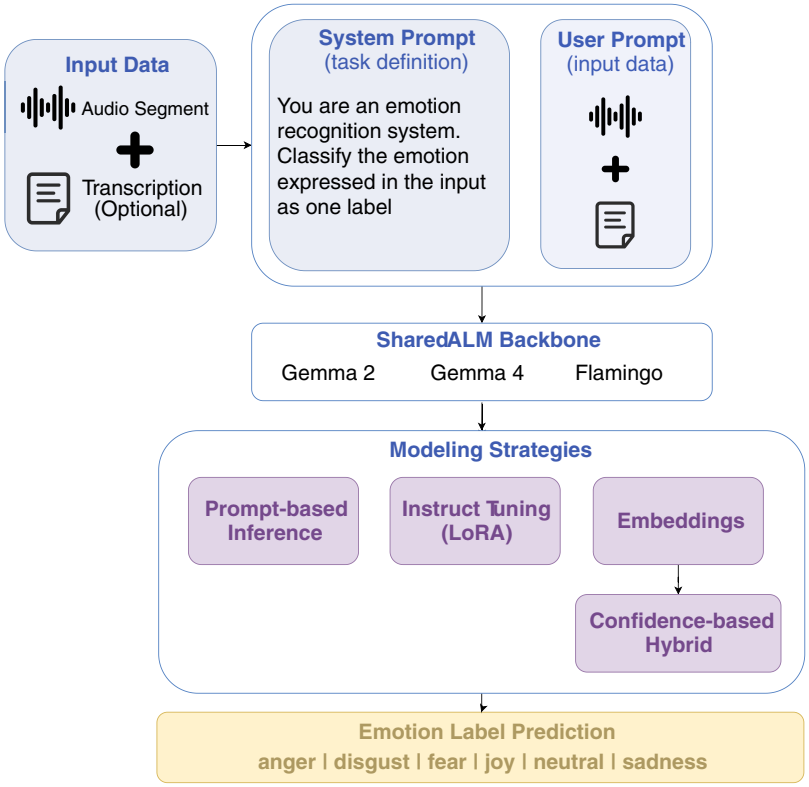


Figure 1: Overview of the proposed system architecture, including the three primary paradigms (prompt-based, instruct-tuned, and embedding-based) and the hybrid routing mechanism.

Table 2: Examples of the input representations used for emotion recognition.

Input Mode	Example Representation
Audio-only	[System]: You are an emotion classifier based on audio. I want you to use the information from the audio to select one of the 6 available emotions. [User]: <audio segment>
Audio + Transcript	[System]: You are an emotion classifier based on audio. I want you to use the information from the audio along with its transcription to select one of the 6 available emotions. [User]: Transcript: "I cannot believe this happened." <audio segment>

3.2 Prompt-Based Architecture

In the prompt-based approach, pretrained ALMs are used directly to perform emotion classification through natural language instructions without requiring task-specific fine-tuning. The prompt construction process is illustrated in Fig. 2. The task is formulated as a conditional generation problem in which the model is prompted to produce the corresponding emotion label based on an input audio segment and, optionally, its transcription.

Following the unified input representation described in Section 3.1, each instance is formatted as a conversational prompt composed of a system message and a user message. In the prompt-based setting, this

format enables the model to receive the task instruction together with the audio signal and, when available, the corresponding transcription. The same structure is also used in the few-shot configuration, where labeled examples are inserted as additional conversational turns before the target instance.

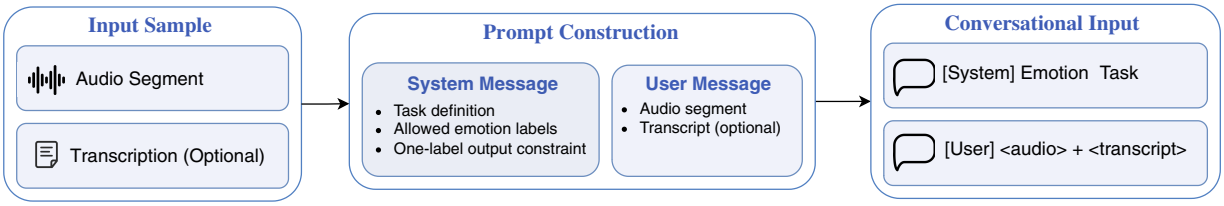


Figure 2: Prompt construction pipeline used to build the conversational input from audio and optional transcript.

Two inference strategies are considered: zero-shot and few-shot prompting (see Fig. 3). In the zero-shot setting, the model is given only the task description and input data. It relies entirely on its pretrained knowledge to infer the correct label. In the few-shot setting, a small number of labeled examples are included in the prompt before the target instance. These examples have the same conversational structure as the target instance and provide additional guidance to the model. Using few-shot prompting allows the model to adapt better to the task distribution without modifying its internal parameters. Conversely, zero-shot inference evaluates the model's inherent generalization capabilities. This comparison allows us to evaluate the effectiveness of prompting strategies for emotion classification in multimodal audio scenarios.

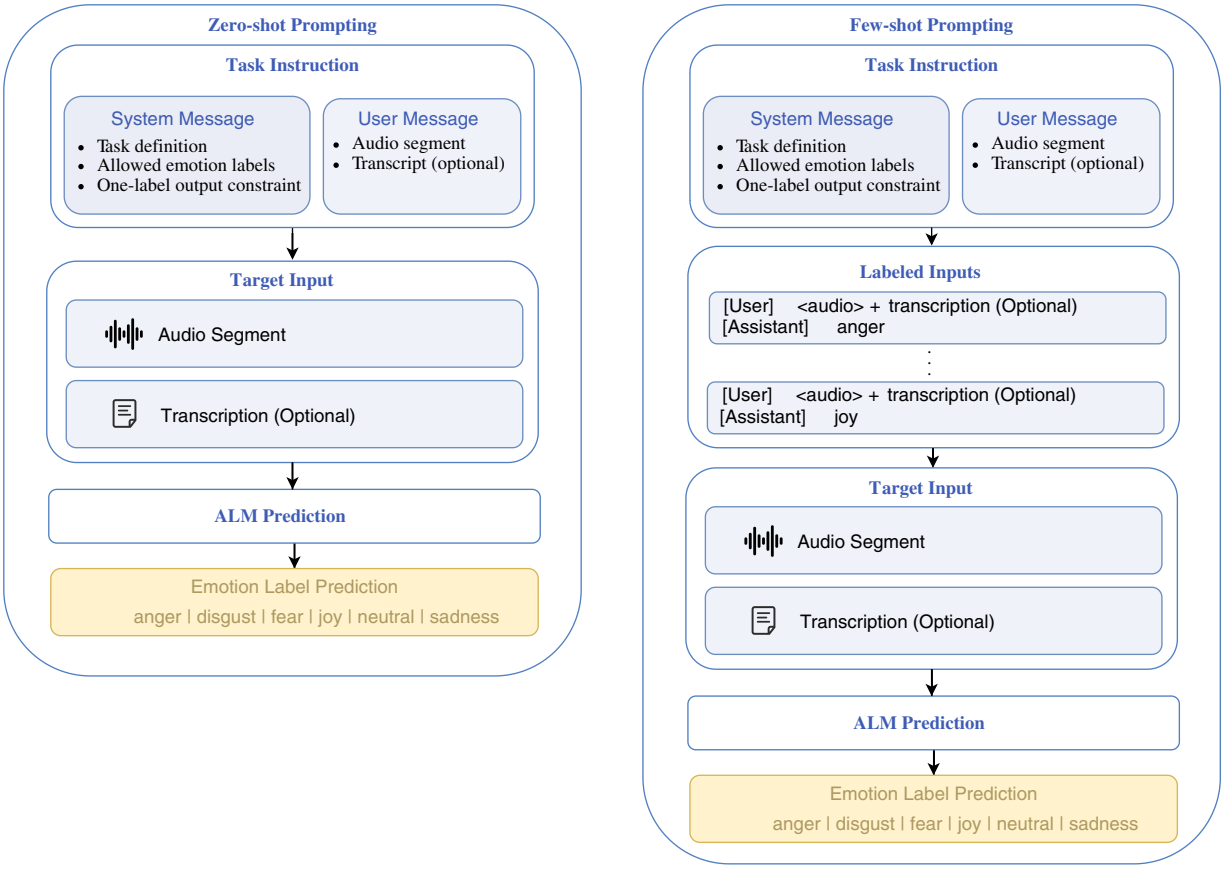


Figure 3: Zero-shot and few-shot prompting architectures and pipelines.

3.3 Instruct-Tuned Models

The second strategy involves adapting ALMs through supervised fine-tuning. This approach aligns the model more closely with the target task by learning from labeled examples in a controlled training setting. To this end, the original dataset is reformulated into a conversational instruction-following format, where each training instance contains a system prompt, a user message with the multimodal input, and an assistant response corresponding to the ground-truth emotion label.

Depending on the selected input mode, the user message includes either the audio signal alone or the audio signal together with its transcription. During training, the model is optimized to generate the emotion label as the assistant response. The generated prediction is compared with the ground-truth label to compute the training loss, and the resulting gradients are used to update only a small set of trainable LoRA parameters while keeping the original pretrained model weights frozen. This strategy enables efficient task adaptation while reducing memory consumption and computational cost, as illustrated in Fig. 4.

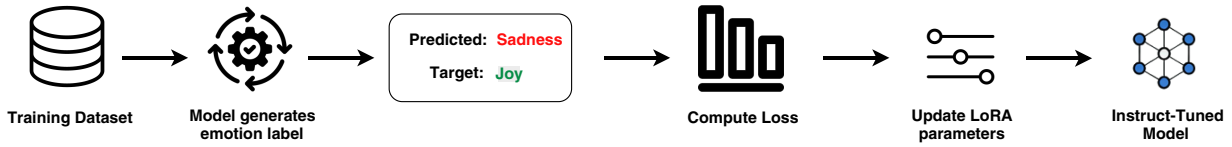


Figure 4: Overview of the instruction-tuning process with LoRA. The training dataset is reformulated into conversational examples, the model generates an emotion label, the prediction is compared with the target label to compute the loss, and only the LoRA parameters are updated while the pretrained model weights remain frozen.

Formally, given a pretrained weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA introduces a low-rank update ΔW such that:

$$W' = W + \Delta W = W + \frac{\alpha}{r} BA, \tag{1}$$

where $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices, r is the rank, and α is a scaling factor. During instruction tuning, the original matrix W remains frozen, and only the LoRA parameters A and B are updated.

3.4 Embedding-Based Architecture

The third paradigm is an embedding-based architecture in which ALMs serve as feature extractors rather than direct predictors. This approach decouples representation learning from the classification stage, allowing for more flexible and computationally efficient modeling. In this approach, input samples are processed using pretrained ALMs to obtain dense vector representations. Each input, consisting of an audio segment and, optionally, its transcription, is formatted according to the conversational structure described in previous sections. The model processes the input and produces hidden states that capture acoustic and semantic information. We then extract the final hidden state corresponding to the last token of the input sequence, which serves as a compact representation of the multimodal input. This embedding serves as a high-level feature vector summarizing the emotional content of the input. Leveraging pretrained ALMs in this manner allows us to benefit from their rich multimodal representations while reducing computational cost at prediction time by eliminating the need for full generative inference.

As shown in Fig. 5, the extracted embeddings are used as input to a lightweight classification model based on an MLP. The classifier is a feed-forward neural network with a single hidden layer, a nonlinear

activation function, and dropout regularization. The classification head is trained using labeled data to predict the emotion category associated with each input. Since the ALM parameters remain fixed during this process, training is significantly more efficient than full model fine-tuning. Additionally, this setup allows for rapid experimentation with different classifier configurations and hyperparameters.

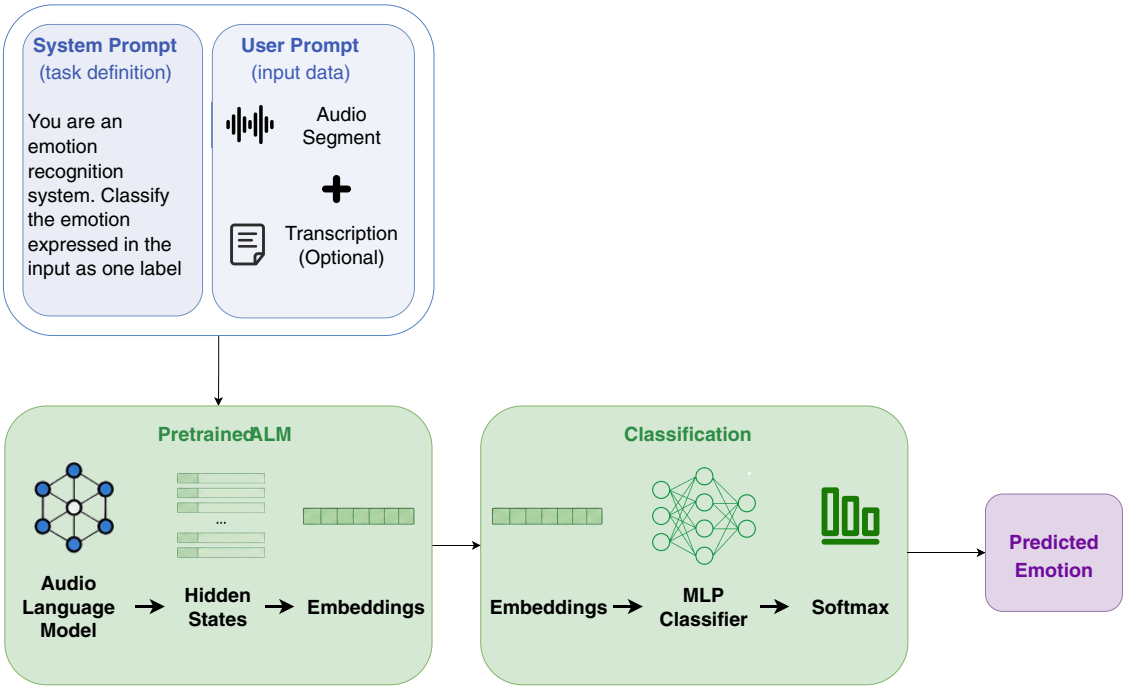


Figure 5: Classification using embeddings extracted from ALMs.

Formally, let x_i denote the conversational multimodal input for sample i, including the audio signal and, when available, its transcription. The frozen ALM f_θ produces a sequence of hidden states $H_i = f_\theta(x_i)$, from which we use the final hidden state $\mathbf{h}_i$ as the input representation for the classifier. The embedding-based prediction is then defined as:

$$\mathbf{h}_i = H_i[T_i], \qquad \mathbf{z}_i = W_2\sigma(W_1\mathbf{h}_i + \mathbf{b}_1) + \mathbf{b}_2,$$
$$\mathbf{p}_i = \mathrm{softmax}(\mathbf{z}_i), \qquad \hat{y}_i^{\mathrm{emb}} = \arg\max_k \mathbf{p}_i(k). \tag{2}$$

where T_i denotes the last position of the hidden-state sequence, σ is the nonlinear activation function, and $\mathbf{p}_i$ represents the probability distribution over the emotion labels.

In addition to extracting embeddings from ALMs, we also include Whisper as a pretrained acoustic baseline. Whisper is not treated as an audio-language model in the prompting or instruction-tuning settings, since it is primarily designed for automatic speech recognition rather than instruction-following emotion classification. Instead, we use the encoder representations extracted from the audio signal and train the same MLP classifier used in the embedding-based experiments. This allows us to compare ALM-derived representations with a strong pretrained speech model under a controlled classification setting.

3.5 Hybrid Architecture

To leverage the strengths of both embedding-based classification and instruction-tuned inference, we propose a hybrid strategy based on confidence-driven routing. This architecture combines the efficiency of

embedding-based classification with the expressive capabilities of the instruction-tuned Flamingo model by applying computationally expensive inference only when necessary. As shown in Fig. 6, the system dynamically routes each input through one of two processing branches.

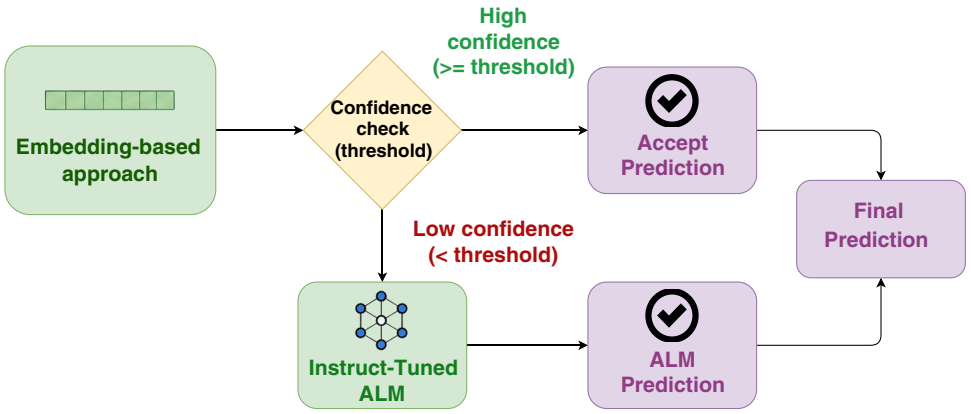


Figure 6: Hybrid architecture pipeline.

In the first stage, the embedding-based pipeline processes the input samples. The pretrained ALM generates a dense representation that is fed into the MLP classifier, which produces a probability distribution over emotion classes. The confidence of each prediction is defined as the maximum probability obtained from the classifier's softmax output. Because they provide confidence proxies through the softmax output of the MLP classifier, embedding-based classifiers are particularly suitable for this routing strategy. This allows for uncertainty-aware decision-making, which is considerably more difficult to achieve in prompt-based generative inference. Predictions with high confidence are accepted as final outputs directly. In contrast, samples with low confidence are considered uncertain and passed to a second stage for further processing. The routing mechanism is controlled by class-specific confidence thresholds. Instead of using a single global threshold for all emotions, we define an individual threshold for each emotion class. This allows the router to account for differences in confidence calibration across emotions, since some classes may require stricter acceptance criteria than others. Lower threshold values favor the embedding-based classifier for a given emotion, improving computational efficiency. Higher values increase reliance on the generative fallback model for that emotion, which may improve robustness in more uncertain cases at the cost of additional computation.

The confidence score used by the router is defined as the maximum probability assigned by the embedding-based classifier. Given the predicted embedding-based label and its associated class-specific threshold, the final hybrid prediction is formalized as:

$$c_i = \max_k \mathbf{p}_i(k), \qquad \hat{y}_i = \begin{cases} \hat{y}_i^{\text{emb}}, & \text{if } c_i \geq \tau_{\hat{y}_i^{\text{emb}}}, \\ \hat{y}_i^{\text{gen}}, & \text{if } c_i < \tau_{\hat{y}_i^{\text{emb}}}, \end{cases} \tag{3}$$

where $\hat{y}_i^{emb}$ is the prediction of the embedding-based classifier and $\hat{y}_i^{gen}$ is the prediction generated by the instruction-tuned Flamingo model. Therefore, the class-specific threshold vector τ controls the trade-off between computational efficiency and reliance on the generative fallback model.

For low-confidence samples, the system uses the best-performing instruct-tuned model, Flamingo, as a fallback mechanism. Unlike the embedding-based stage, which performs lightweight classification over fixed representations, the instruct-tuned Flamingo model performs full generative inference based on the

multimodal input. This design choice is motivated by the strong dependence of hybrid systems on the quality of the fallback component. Because Flamingo demonstrated the best overall performance among all evaluated instruction-tuned models, it was chosen as the shared fallback model for all hybrid configurations.

The final prediction is determined by combining the outputs from both stages. High-confidence predictions from the embedding-based classifier are retained, and low-confidence cases are replaced with outputs from the generative model. This hybrid strategy offers a flexible, adaptive framework integrating fast, embedding-based classification with a robust, instruction-tuned fallback model. The goal is to enhance robustness for uncertain samples while maintaining computational efficiency for high-confidence predictions.

3.6 Supported Models

To comprehensively evaluate the proposed architectures, we use a variety of pretrained ALMs with different design characteristics and capabilities. These models vary in size, training objectives, and multimodal integration strategies. This allows us to analyze the behavior of each approach across different architectures. The selected models include recent instruction-tuned and multimodal generative systems that can process audio inputs. These models include Gemma-based models and Audio Flamingo variants (see Table 3). These models provide native support for multimodal inputs, enabling the integration of audio signals and textual prompts within a unified framework. All models are used under a consistent input representation and prompting interface to ensure comparability across experiments. Depending on the architectural paradigm, the models are used as either generative predictors, fine-tuned instruction followers, or feature extractors for embedding-based classification. This diverse selection of ALMs enables us to systematically evaluate the effectiveness of various modeling strategies and assess the impact of architectural choices on the performance of multimodal emotion classification.

Table 3: ALMs evaluated in this study for SER, including the model aliases used throughout the paper, their corresponding pretrained checkpoints, model families, and approximate number of parameters (in billions).

Alias	Model	Family	Size (B Parameters)
Gemma2	google/gemma-3n-E2B-it [55]	Gemma	2
Gemma4	google/gemma-3n-E4B-it [55]	Gemma	4
Flamingo	nvidia/audio-flamingo-next-hf [7]	AF-Next	8

4 Experimental Setup

This section describes the dataset, evaluation protocol, implementation details, and experimental configurations used to evaluate the various ALM-based approaches.

4.1 Dataset

We conducted our experiments using the Spanish MEACorpus dataset [12], a multimodal corpus designed for emotion recognition in real-world conditions. The dataset consists of audio segments extracted from natural interactions, each of which is annotated with one of the following six emotion categories: anger, disgust, fear, joy, neutral, and sadness. Each sample includes an audio signal and its corresponding textual transcription, enabling unimodal (audio-only) and multimodal (audio + text) configurations. This allows us to evaluate the impact of incorporating linguistic information alongside acoustic cues. For training-based approaches, we adhere to the original dataset split and partition the training set into two subsets. Specifically,

10% of the training data is used as a validation set for hyperparameter tuning and model selection, and the remaining 90% is used for training. The emotion and split distributions are reported in Table 4.

Table 4: Emotion and split distribution of the MEACorpus dataset.

Emotion	Train	Test	Total
anger	534 (13.02%)	134 (13.05%)	668 (13.02%)
disgust	960 (23.40%)	240 (23.37%)	1200 (23.40%)
fear	29 (00.71%)	8 (00.78%)	37 (00.72%)
joy	514 (12.53%)	129 (12.56%)	643 (12.54%)
neutral	1600 (39.01%)	399 (38.85%)	1999 (38.97%)
sadness	465 (11.34%)	117 (11.39%)	582 (11.35%)
Total	4102 (79.98%)	1027 (20.02%)	5129

We evaluate model performance using precision, recall, and F1-score. For each emotion class c, precision and recall are defined as:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \qquad R_c = \frac{TP_c}{TP_c + FN_c}, \tag{4}$$

where TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively. The F1-score for each class is computed as the harmonic mean of precision and recall:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{5}$$

Given the set of emotion classes C, we report both macro-averaged and weighted F1-scores. The macro F1-score assigns equal importance to all emotion categories and is computed as:

$$F1_{\text{macro}} = \frac{1}{|C|}\sum_{c\in C} F1_c. \tag{6}$$

In contrast, the weighted F1-score accounts for class imbalance by weighting each class according to its support n_c:

$$F1_{\text{weighted}} = \sum_{c\in C}\frac{n_c}{N}F1_c, \tag{7}$$

where N is the total number of samples.

Macro F1 is particularly relevant in this work because the dataset is imbalanced and minority emotions should contribute equally to the final evaluation. Weighted F1 is also reported to provide an overall measure that reflects the class distribution. Additionally, we use normalized confusion matrices to analyze class-specific behavior and identify common misclassification patterns. Finally, we compare our results with those of the best-performing method reported in the original MEACorpus study.

4.2 Prompt-Based Experiments

We evaluate both zero-shot and few-shot settings for prompt-based inference. In the zero-shot configuration, the model is prompted with only a task description and the input sample. In contrast, the few-shot

setting adds a few labeled examples to the prompt, allowing the model to better align with the target classification task.

Two configurations are considered regarding the input representation. In the audio-only setting, the model receives only the raw audio signal. In the audio + transcript setting, the input is enriched with the corresponding textual transcription. This allows the model to leverage both acoustic and linguistic information.

For few-shot learning, we set k to six examples, selecting one instance per emotion class to ensure a balanced representation of all categories. These examples were randomly sampled from the training set using a fixed seed under the constraint of selecting one example for each emotion label. These demonstrations were incorporated into the prompt as additional conversational turns preceding the target query. This approach aligns with the standard in-context learning paradigm, wherein the model infers the task from the demonstration examples.

All models use deterministic short-output generation to produce emotion predictions, as shown in Listing 4.1. In practice, generation is limited to a small number of tokens and performed using greedy decoding. The generated text is then normalized to one of the predefined emotion labels whenever possible. This procedure reduces variability in the outputs and ensures more consistent and comparable predictions across models. Interaction with the model follows a conversational structure consisting of a system prompt that defines the task and a user message including the input (audio and optional transcript). In the few-shot setting, additional exchanges between the user and assistant are inserted before the final query to provide labeled examples, maintaining a unified prompting format across all evaluated models. An example of the prompting format is shown below:

Listing 1: Example of the conversational prompt format used in zero-shot and few-shot emotion classification. In the few-shot setting, additional user-assistant exchanges containing labeled examples are inserted before the final query.

```
System:
You are an audio emotion classification system.
Classify the emotion expressed in the input audio.
Use both vocal cues and semantic content when available.
Respond with exactly one label from the predefined set.
User:
Classify the emotion of the following audio (and transcript if
provided).
Allowed labels: {anger, disgust, fear, joy, neutral, sadness}.
Assistant:
sadness.
```

4.3 Instruct-Tuning Setup

For instruction tuning, we fine-tune pretrained ALMs using a conversational reformulation of the dataset. Each sample is represented as a sequence of system, user, and assistant messages. The assistant response corresponds to the ground-truth emotion label. This formulation allows the model to learn the relationship between multimodal inputs (audio and optional transcripts) and specific emotion categories within an instruction-following framework.

We employ parameter-efficient fine-tuning with LoRA, which updates only a subset of the model parameters while maintaining a fixed backbone. The training process uses a supervised fine-tuning strategy that optimizes the cross-entropy loss over the generated tokens. We perform model selection based on validation loss using a development split derived from the training data. We conduct training using the AdamW optimizer with a learning rate of 2×10^{-4} for a total of three epochs. Due to the variable length of audio inputs and the resulting differences in sequence size after multimodal tokenization, we use a per-device batch size of one, which avoids excessive padding and reduces memory overhead. This is particularly relevant when processing long audio sequences. To mitigate the instability associated with small batch sizes, we apply gradient accumulation over four steps, effectively simulating a larger batch size of four. This allows for more stable gradient estimates while remaining within hardware constraints.

More specifically, the LoRA adapters were configured with rank $r = 16$, scaling factor $\alpha = 32$, and dropout $p = 0.05$. Bias parameters were not updated. The adapters were inserted into both the attention projection layers and the feed-forward projection layers, targeting `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The task type was set to causal language modeling, since the model was optimized to generate the emotion label as the assistant response in the conversational formulation.

4.4 Embedding-Based Experiments

In the embedding-based approach, pretrained ALMs serve as feature extractors, providing fixed-size representations of each input sample. These embeddings are then used to train a lightweight MLP classifier. The classification head is a feed-forward neural network with one hidden layer, ReLU activation, and dropout regularization.

To optimize the classifier, we perform a grid search over multiple hyperparameters, exploring learning rates in $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}, 2 \times 10^{-5}\}$, weight decay values in $\{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}\}$, training epochs in $\{20, 30, 50\}$, batch sizes in $\{16, 32\}$, and dropout rates in $\{0.10, 0.15, 0.20, 0.25\}$. We apply early stopping with a patience of 10 based on the validation macro F1-score to prevent overfitting. To improve efficiency, embeddings are cached and reused across experiments to avoid redundant computations. Table 5 shows the optimized hyperparameter configuration for each model after performing the 480 possible combinations. These configurations were used to generate the final MLP predictions.

Table 5: Best hyperparameter configurations obtained through grid search for the MLP classifier trained on ALM embeddings. The selected values correspond to the settings that achieved the highest macro F1-score on the development set for each evaluated model.

Model	Learning Rate	Weight Decay	Epochs	Batch Size	Dropout
Gemma2 [55]	0.001	0.00001	50	32	0.15
Gemma4 [55]	0.001	0.001	50	16	0.2
Flamingo [7]	0.0005	0.00001	50	32	0.1

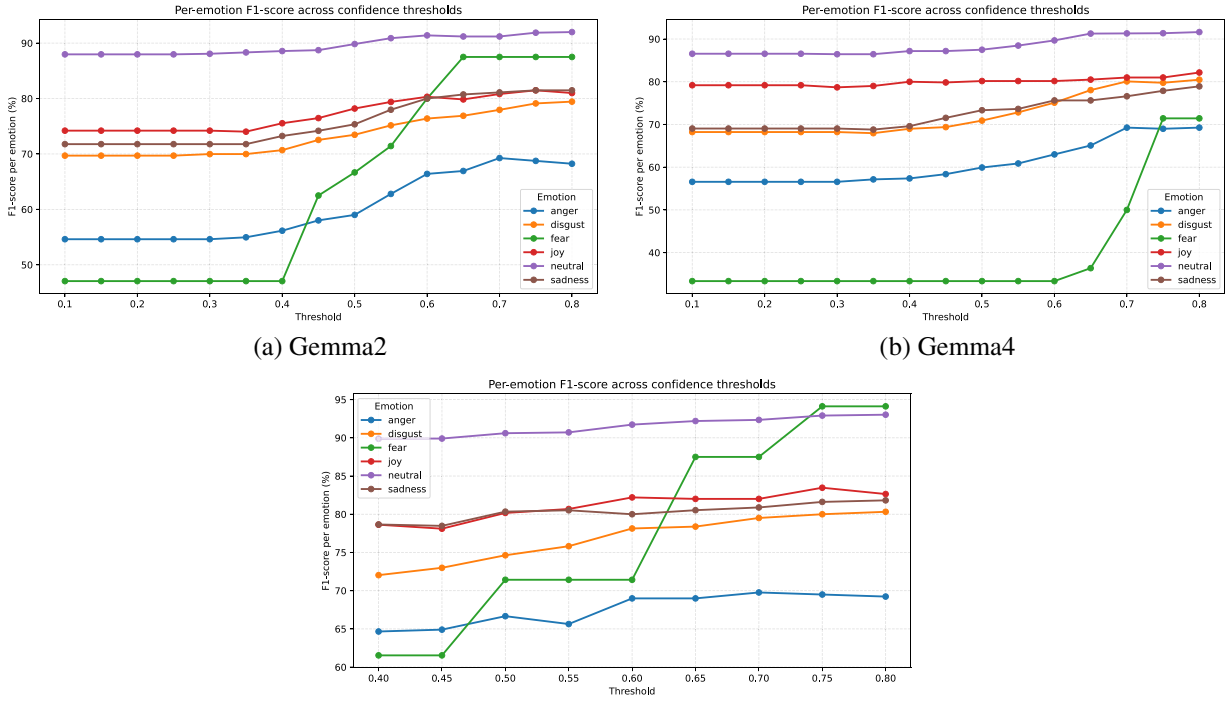
For the Whisper baseline, we follow the same embedding-based classification protocol. However, since Whisper does not use a conversational audio-language interface, the audio signal is processed directly by the Whisper encoder. The resulting hidden representations are aggregated into a fixed-size vector and used as input to the MLP classifier. This configuration is evaluated only in the audio-only setting and serves as a pretrained acoustic baseline rather than as an ALM-based prompting or instruction-tuning approach.

4.5 Hybrid Approach Configuration

In the hybrid architecture, we combine an embedding-based classifier with a generative ALM through a confidence-based routing mechanism.

The embedding-based component uses the same hyperparameter configuration described in Section 4.4. The fallback component relies on the instruct-tuned Flamingo model, which is described in Section 4.3 and achieved the best performance among the instruction-tuned configurations. Predictions from the embedding-based model are then associated with a confidence score derived from the maximum softmax probability. We define a class-specific threshold vector $\boldsymbol{\tau}$, where each component corresponds to one emotion class. Samples whose confidence score is greater than or equal to the threshold associated with the predicted emotion are accepted directly. Samples whose confidence score is below that class-specific threshold are considered low-confidence predictions and are reprocessed using the generative model.

We evaluate three hybrid configurations using embeddings extracted from Gemma2, Gemma4, and Flamingo, respectively. Low-confidence predictions are selectively reprocessed using the instruct-tuned Flamingo model as a shared fallback mechanism in all cases. For threshold selection, we perform a class-wise search over the validation split. Instead of selecting a single global confidence threshold, we optimize an independent threshold for each emotion class. Candidate thresholds are evaluated within the ranges shown in Fig. 7: from 0.1 to 0.8 for Gemma2 and Gemma4, and from 0.4 to 0.8 for Flamingo. For each class, the selected threshold is the value that maximizes the class-specific F1-score on the validation set. The resulting threshold vector is then fixed and applied to the test set. The selected class-specific thresholds, together with the proportion of test samples routed to the Flamingo fallback model, are reported in Table 6. This strategy allows the hybrid router to adapt to the different confidence distributions of each emotion class, rather than applying the same acceptance criterion to all predictions.



(a) Gemma2

(b) Gemma4

(c) Flamingo

Figure 7: Class-wise F1-score obtained across confidence thresholds for the hybrid routing mechanism using embeddings extracted from Gemma2, Gemma4, and Flamingo. Each curve represents one emotion class.

Table 6: Class-specific confidence thresholds selected on the validation split for each hybrid configuration. The last column reports the proportion of test samples whose embedding-based confidence was below the corresponding class-specific threshold and were therefore routed to the instruct-tuned Flamingo fallback model.

Embedding Backbone	Anger	Disgust	Fear	Joy	Neutral	Sadness	Fallback (%)
Gemma2	0.70	0.80	0.65	0.75	0.80	0.75	52.68%
Gemma4	0.70	0.80	0.75	0.80	0.80	0.80	57.84%
Flamingo	0.70	0.80	0.75	0.75	0.80	0.80	47.13%

To further illustrate the threshold-selection process, Fig. 7 shows the class-wise F1-score obtained across different confidence thresholds for each embedding backbone. The curves reveal that the optimal threshold is not uniform across emotions. Some classes, such as neutral, remain relatively stable across the explored range, whereas others, especially fear, show a stronger dependence on the selected threshold. This behavior supports the use of class-specific thresholds instead of a single global threshold for all emotion categories.

5 Results and Discussion

This section presents a detailed analysis of the evaluation results obtained by the different approaches explored in this work. All models were evaluated on the test split, following the experimental protocol described in Section 4.

The results presented in Tables 7 and 8 reveal a clear performance hierarchy among the evaluated ALM-based strategies. Prompt-based approaches obtained the lowest scores, with the best few-shot configuration achieving a macro F1 score of 42.79%. Embedding-based classification substantially improves upon these results, achieving a macro F1-score of 72.61% while providing a computationally efficient alternative to generative methods. The proposed hybrid approach boosts performance even further, reaching 83.55% macro F1 and 84.37% weighted F1 by combining embedding-based predictions with selective generative inference. Instruction-tuned Flamingo remains highly competitive, achieving 83.32% macro F1 and 84.07% weighted F1. These results suggest that task-specific supervised adaptation is essential for utilizing ALMs in SER, while class-specific confidence-driven routing can further refine the final predictions.

The results obtained with prompt-based approaches reveal several limitations when applying ALMs directly to emotion classification tasks. In the zero-shot setting, all models perform relatively poorly, with macro F1 scores ranging from 33% to 40%. Incorporating transcription information does not consistently improve performance across all models and settings. While it provides noticeable gains in some few-shot configurations, its effect is marginal or even negative in several zero-shot cases. For example, in the Flamingo zero-shot setting, adding the transcript slightly reduces the macro F1-score, whereas in the few-shot setting it produces only a minor improvement. This behavior may be attributed to the difficulty of incorporating multiple multimodal examples within the context window and the sensitivity of these models to prompt design and example selection. Overall, these findings suggest that prompting alone is insufficient for achieving competitive performance in this task.

In contrast, instruction tuning substantially improves all models. The fine-tuned Flamingo model achieves the best result among the instruction-tuned configurations, with a macro F1-score of 83.32%, outperforming its zero-shot counterpart by over 43 percentage points. Similar gains are observed for Gemma2 and Gemma4. These improvements are due to the models being explicitly aligned with the target task through supervised training. This enables the models to better capture the relationship between multimodal inputs and emotion labels. These results underscore the importance of task-specific adaptation when applying ALMs to structured classification problems.

Table 7: Macro-averaged precision (P), recall (R), and F1-score (%) obtained by each evaluation strategy and model on the MEACorpus test set. The highest macro F1-score within each prompt-based configuration is highlighted in bold.

Strategy	Gemma2 [55]			Gemma4 [55]			Flamingo [7]		
	P	R	F1	P	R	F1	P	R	F1
Prompt-based									
Zero-Shot	40.8671	39.9328	33.3305	37.3457	46.9812	37.8550	56.4454	41.5477	**39.7055**
Zero-Shot (w/trans.)	42.2516	41.8190	36.4471	37.5194	44.2820	36.9908	53.9379	39.9130	**38.7054**
Few-Shot	37.1968	27.4083	18.0238	35.3040	39.2298	**32.5518**	18.3861	19.2754	14.7744
Few-Shot (w/trans.)	46.6431	48.0814	38.4107	44.8810	52.9959	**42.7883**	20.3811	19.7203	15.2733
Instruct-Tuning									
Instruct-Tuning	49.0940	55.3136	50.3767	68.2039	74.5951	65.1186	84.5785	82.8033	**83.3190**
Embedding-based									
Embedding-Based	63.3926	62.7073	60.4420	65.2481	59.7911	60.7000	80.2630	69.1905	**72.6083**
Hybrid									
Hybrid	83.4768	80.5605	81.8282	81.9956	76.4294	78.8210	84.1624	83.3630	**83.5516**

Table 8: Comparison of the best-performing configuration of each strategy using macro-averaged and weighted precision (P), recall (R), and F1-score (%) on the MEACorpus test set. The highest precision, recall and macro F1-score within each prompt-based configuration is highlighted in bold.

Strategy	Macro Avg			Weighted Avg		
	P	R	F1	P	R	F1
Few-Shot	44.8810	52.9959	42.7883	51.8593	49.6592	49.3796
Tuning	**84.5785**	82.8033	83.3190	84.4655	84.2259	84.0713
Embeddings	80.2630	69.1905	72.6083	81.6322	81.0127	80.2956
Hybrid	84.1624	**83.3630**	**83.5516**	**84.5074**	**84.5180**	**84.3696**

The embedding-based approach delivers strong, consistent performance across models and clearly outperforms all prompt-based configurations. With Flamingo, this method achieves up to 72.61% macro F1, demonstrating that pretrained representations extracted from ALMs are sufficiently expressive for emotion classification when combined with a lightweight classifier. Furthermore, this approach is more computationally efficient than generative methods because it avoids autoregressive decoding and relies on a simple feed-forward classification head.

Finally, the hybrid approach shows consistent improvement over embedding-based classification across all evaluated backbones. For the Flamingo-based configuration, the hybrid approach slightly surpasses the standalone instruction-tuned model in macro F1-score, increasing it from 83.32% to 83.55%. Although the hybrid model does not improve macro precision over instruction tuning, it obtains the highest weighted F1-score, reaching 84.37%. This suggests that the confidence-based routing mechanism improves the final decision process by selectively replacing uncertain embedding-based predictions with outputs from the instruction-tuned fallback model. As shown in Table 8, these results suggest that class-specific confidence-based routing can improve overall classification performance, although the gain over instruction tuning remains small.

In addition to the ALM-based strategies, we include Whisper as a pretrained acoustic baseline to contextualize the quality of the representations learned by audio-language models. Whisper is not evaluated under prompting or instruction-tuning settings because it is primarily designed as an automatic speech recognition model rather than an instruction-following audio-language model. Instead, we use its encoder representations as input to the same MLP classifier used in the embedding-based experiments, allowing a controlled comparison between ALM-derived embeddings and a strong pretrained acoustic representation.

The Whisper encoder baseline achieves a macro F1-score of 79.49% and a weighted F1-score of 82.70%. This result is substantially higher than the prompt-based configurations and also outperforms the standalone embedding-based results obtained with the evaluated ALMs. However, it remains below the best instruction-tuned ALM configuration, which reaches 83.32% macro F1 and 84.07% weighted F1, and below the proposed hybrid approach, which achieves the highest overall scores with 83.55% macro F1 and 84.37% weighted F1. These results indicate that pretrained acoustic representations are highly competitive for Spanish SER, but that supervised task alignment and confidence-driven routing provide additional gains in our experimental setting.

When compared with the original MEACorpus baseline, the best ALM-based configuration remains below the specialized task-specific system, for which an 87.74% macro F1-score was reported. This gap suggests that, although general-purpose ALMs provide a flexible and competitive framework for speech emotion recognition, specialized acoustic models designed explicitly for SER continue to offer stronger performance under this benchmark.

To assess the robustness of the internally evaluated configurations, we performed non-parametric bootstrap resampling on the test set with 10,000 iterations. For each bootstrap sample, we recomputed macro and weighted F1-scores and estimated 95% confidence intervals.

The bootstrap analysis shows that instruction tuning and the hybrid approach obtain very similar performance. Although the hybrid approach achieves a slightly higher macro F1-score and weighted F1-score, the confidence intervals of both methods are highly overlapping, indicating that their performance is comparable. In contrast, both instruction tuning and the hybrid approach clearly outperform the embedding-based classifier in macro F1, with differences above 10 percentage points. The corresponding 95% bootstrap confidence intervals are reported in Table 9.

Table 9: Bootstrap 95% confidence intervals for the best-performing configuration of each strategy on the MEACorpus test set.

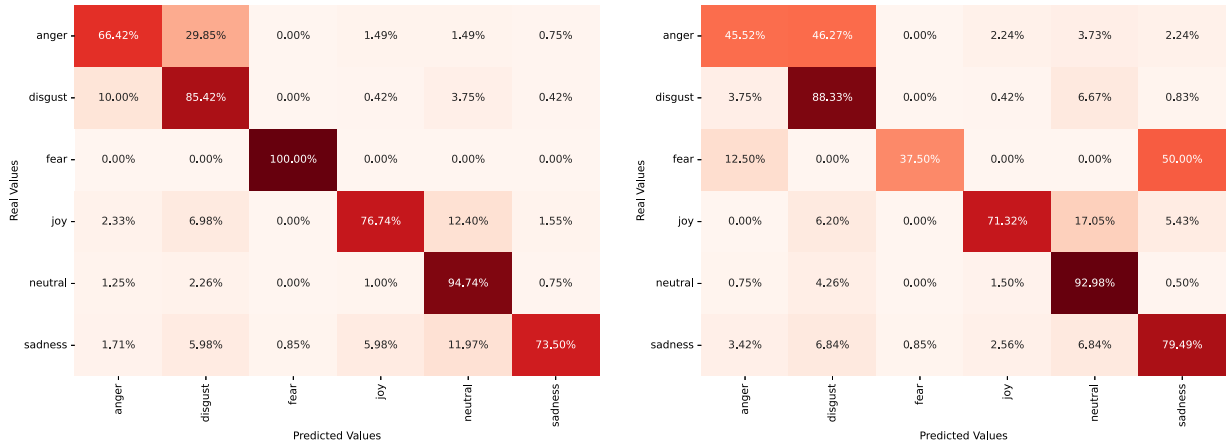
Strategy	Macro F1 (95% CI)	Weighted F1 (95% CI)
Few-Shot	42.79 [39.40, 46.16]	49.38 [46.18, 52.68]
Instruct-Tuning	83.32 [79.50, 86.09]	84.07 [81.75, 86.30]
Embeddings	72.61 [64.17, 78.49]	80.30 [77.66, 82.85]
Hybrid	83.55 [79.81, 86.29]	84.37 [82.04, 86.61]

The hybrid approach obtains the highest point estimate in terms of both macro F1 and weighted F1, while instruction tuning remains highly competitive, with strongly overlapping confidence intervals. The embedding-based approach remains a competitive and computationally efficient alternative, although it is clearly outperformed by both instruction tuning and hybrid routing. Prompt-based methods, on the other hand, yield substantially lower scores. These findings demonstrate that supervised adaptation is essential for maximizing the capabilities of audio-language models in speech emotion recognition, while class-specific confidence-based routing can further refine the final predictions and provide the best overall point estimate.

5.1 Error Analysis

To further analyze model behavior, we examine the normalized confusion matrices of representative approaches. Our focus is on Flamingo-based models because Flamingo consistently achieves the strongest performance across all evaluated paradigms. This makes Flamingo a suitable reference for comparing different strategies under the same backbone.

Fig. 8 compares the confusion matrices of the instruct-tuned Flamingo model (Fig. 8a) and the embedding-based Flamingo model (Fig. 8b). The instruction-tuned model performs well across most emotion classes, particularly for neutral and disgust, which are correctly classified in 94.74% and 85.42% of cases, respectively. However, there is some confusion between anger and disgust, suggesting that these emotions share similar acoustic or semantic patterns that are difficult to distinguish. In contrast, the embedding-based model exhibits a more balanced distribution of predictions across classes. However, it shows increased confusion between semantically related emotions, such as sadness, neutral, and fear, indicating a more general but less discriminative representation. The prompt-based approach shows clear performance degradation, with strong biases toward dominant classes, such as neutral, and particularly low accuracy for anger and disgust. These observations confirm that prompt-based methods are less effective for this task, whereas instruct-tuned models achieve better class separation and more reliable predictions.



(a) Flamingo with instruct-tuning **(b) Flamingo embeddings**

Figure 8: Normalized confusion matrices for Flamingo-based approaches. **Left**: instruct-tuned model. **Right**: embedding-based model.

As shown in Fig. 9, the hybrid configuration exhibits more conservative prediction behavior than the standalone instruct-tuned model. The confidence-based routing mechanism slightly improves several class-level results, particularly for neutral, sadness, anger, and joy, although performance for disgust decreases and fear remains unchanged. This behavior helps explain why the hybrid approach achieves the highest overall macro F1-score, while still preserving some class-specific confusion patterns. Specifically, the hybrid configuration inherits the robustness of the instruct-tuned fallback model while retaining some limitations of the embedding-based classifier. This suggests that the fallback mechanism improves difficult and low-confidence predictions, yet does not entirely eliminate the structural biases introduced during the initial embedding-based classification stage.

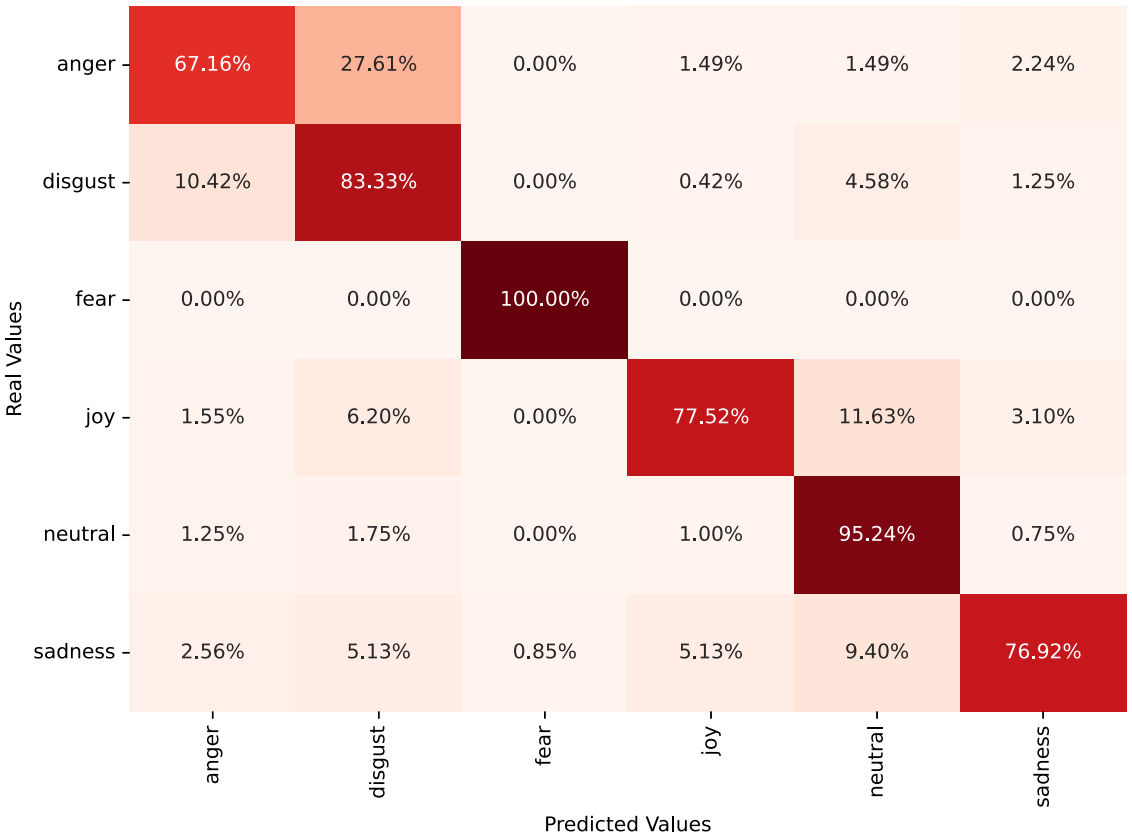


Figure 9: Normalized confusion matrix for Flamingo-Hybrid approach.

Table 10 provides a qualitative comparison of representative examples across different modeling paradigms. These examples highlight several important patterns. First, prompt-based methods tend to produce unstable predictions. They often misclassify emotionally charged inputs as neutral or assign incorrect emotion categories, as observed in Error Type 1. Second, while the embedding-based approach is more stable, it struggles with fine-grained distinctions between emotions, particularly in subtle or ambiguous cases. In contrast, the instruct-tuned model better captures these nuances, as shown in Error Type 2. Finally, Error Type 3 shows that some samples are difficult for all approaches and usually involve complex or ambiguous emotional cues. Prompt-based methods exhibit the highest proportion of unstable or incorrect predictions. Embedding-based models reduce this effect, but still struggle with subtle distinctions. In contrast, instruct-tuned models demonstrate a lower proportion of critical misclassifications, especially in emotionally nuanced cases, suggesting enhanced robustness.

Overall, the results of both the quantitative and qualitative analyses reveal a clear hierarchy among the evaluated approaches. Prompt-based methods are unreliable, whereas embedding-based models provide a stronger and more stable baseline. Instruction-tuned models offer robust class separation and substantially improve over prompting and embedding-based classification. Finally, the hybrid strategy provides the best overall performance by combining the robustness of the instruction-tuned fallback model with the efficiency of embedding-based predictions.

Table 10: Qualitative error analysis comparing different modeling approaches.

Text	Truth	Prompting	Embeddings	Instruct
Error Type 1: Prompting fails, while other approaches succeed				
Después de haber tenido pufos en los test, pufos en las mascarillas, después de haber dicho que no eran necesarias para ir por la calle, y ahora no las quieren meter hasta en la playa.	Anger	Disgust	Anger	Anger
Mikhail Gorbachev ha muerto. El último presidente de la URSS y padre de la perestroika falleció este martes a la edad de 91 años. Falleció tras una larga y grave enfermedad, dijeron fuentes del Hospital Clínico Central de Moscú.	Neutral	Sadness	Neutral	Neutral
Error Type 2: Embeddings fail, instruct-tuning succeeds				
De usar y deteriorar las instituciones del Estado. ¿Está usted con las víctimas o con los verdugos? Nosotros estamos con las víctimas. ¿Y usted? Si hasta de cónomos dice	Anger	Neutral	Disgust	Anger
A las 11 de la mañana hubiera dicho acepto la mano tendida de Pablo Casado, me equivoqué, montemos una ley de pandemias entre los dos grandes partidos y trabajemos codo con codo para luchar contra el virus.	Anger	Neutral	Neutral	Anger
Error Type 3: All approaches fail				
Frente a las 28.000 víctimas que reconoce el Gobierno, el Instituto Nacional de Estadística, el Instituto Carlos III y la patronal a las funerarias coinciden en que entre marzo y junio murieron en España unas 45.000 personas más de lo previsto.	Sadness	Neutral	Neutral	Disgust
Hoy quedan 34 días para que Isabel Díaz Ayuso y el Partido Popular ganen las elecciones de la Comunidad de Madrid. Hoy quedan 34 días.	Joy	Neutral	Disgust	Disgust

5.2 Comparison with MEACorpus Baseline

To provide context for our results, we compare the best-performing configuration in our study with the results reported in the original MEACorpus study. Our best configuration, based on the Flamingo hybrid model with class-specific confidence-driven routing, achieves a macro F1-score of 83.55%, compared with the 87.74% reported by the MEACorpus baseline. Although the gap is moderate, it can be attributed, first, to important differences in modeling assumptions. The MEACorpus baseline relies on task-specific acoustic representations and architectures explicitly designed for speech emotion recognition, whereas the models evaluated in this work are general-purpose audio-language models. While instruction tuning substantially improves task alignment and hybrid routing further refines the final predictions, these strategies may not

fully compensate for the lack of task-specific acoustic inductive biases needed to capture fine-grained paralinguistic cues such as pitch variation, intensity, rhythm, voice quality, or subtle prosodic patterns.

Second, the size and distribution of the dataset also affect the adaptation of large ALMs. This limitation is particularly relevant for minority classes. For example, the fear class contains only 29 training samples and 8 test samples, representing less than 1% of the dataset. Since macro F1 assigns equal weight to all emotion categories, poor performance on such underrepresented classes can strongly affect the final score. The evaluation language may also contribute to this gap. Although Spanish is a high-resource language in general NLP, resources specifically designed for speech emotion recognition in Spanish remain more limited than those available for English [12,67]. This limitation affects both supervised adaptation and the evaluation of general-purpose ALMs, since emotional expression depends not only on lexical content but also on language-specific prosodic patterns, intonation, speaking style, and cultural conventions. Moreover, the limited availability of large-scale Spanish SER data may restrict the ability of ALMs to fully adapt to fine-grained emotional cues in this language.

Third, the results suggest that the main limitation is not only the amount of training data, but also the difficulty of distinguishing acoustically and semantically similar emotions. As shown in the error analysis, the best-performing models still exhibit confusion between anger and disgust, and between neutral, sadness, and fear in some configurations. These confusions indicate that ALMs can capture broad emotional patterns after instruction tuning, but they still struggle with fine-grained emotional boundaries when the acoustic or semantic cues are ambiguous.

Finally, overfitting is a relevant concern when adapting large models to relatively small datasets. To reduce this risk, we used parameter-efficient fine-tuning with LoRA, kept the backbone fixed, and selected the best checkpoint according to validation loss. Nevertheless, the limited number of labeled examples, especially for minority emotions, may still constrain generalization. This suggests that further gains may require either larger and more balanced Spanish SER datasets, stronger acoustic pretraining, or more advanced hybrid architectures that better integrate task-specific speech representations with the reasoning capabilities of ALMs.

The results should be interpreted in the context of the Spanish language. Most large-scale audio-language models are trained using multilingual corpora in which English is often dominant. Consequently, the amount of Spanish emotional speech available during pretraining may be substantially lower than the amount of English speech, which could limit the models' ability to capture language-specific emotional and prosodic patterns. Additionally, Spanish SER resources are comparatively scarce, which reduces the availability of task-specific data for adaptation. These factors may explain why general-purpose ALMs lag behind specialized models trained for specific SER benchmarks, such as MEACorpus.

Overall, this comparison shows that general-purpose ALMs are a promising and flexible alternative for Spanish speech emotion recognition, but they do not yet outperform highly specialized acoustic models. The results position ALMs as competitive systems with strong potential for multimodal and instruction-following scenarios, while also highlighting the importance of task-specific acoustic modeling for achieving state-of-the-art SER performance.

6 Conclusion and Future Work

In this study, we evaluated several ALM paradigms for SER, including prompt-based inference, instruction tuning, embedding-based classification, and a hybrid approach. This work provides a systematic analysis of how these strategies perform within a unified experimental framework. Our results show that the hybrid approach with class-specific confidence-driven routing achieves the best overall performance. These findings demonstrate the importance of adapting models to specific tasks and shows that confidence-driven

routing can further refine the final predictions. Nevertheless, the best ALM-based configuration remains below the specialized MEACorpus baseline, achieving 83.55% macro F1-score, compared with 87.74%. This result indicates that general-purpose ALMs are highly competitive, but they do not yet surpass task-specific acoustic models designed explicitly for speech emotion recognition. The Whisper encoder baseline further supports this observation, showing that pretrained acoustic representations are highly competitive but still do not surpass the best instruction-tuned ALM configuration, the best hybrid ALM configuration, or the specialized MEACorpus baseline in our setting. Although prompt-based approaches are flexible and require no additional training, they remain not competitive for this task and are less reliable, especially for distinguishing nuanced emotions. Embedding-based methods offer a more efficient alternative with lower computational costs. They achieve reasonable performance and enable faster inference. Finally, the proposed hybrid strategy combines the efficiency of embedding-based classification with the robustness of instruction-tuned inference, achieving the highest macro and weighted F1-scores among the evaluated ALM-based configurations. The results also suggest that the effectiveness of confidence-driven routing depends heavily on the quality of the fallback model. In our experiments, the Flamingo-based hybrid configuration improved the handling of low-confidence predictions through the use of the instruct-tuned Flamingo fallback model. Beyond performance comparisons, our findings suggest that ALMs can capture useful emotional cues from audio signals when they are adapted through instruction tuning or used as representation extractors.

Despite these promising results, several limitations must be acknowledged. First, evaluating the models on a single dataset may limit the generalizability of the conclusions. Second, prompt-based approaches are sensitive to prompt design and may produce variable outputs. Third, embedding-based methods rely on fixed representations that may not fully capture task-specific nuances. Future work will focus on several areas. First, we plan to explore more advanced prompting strategies, including multi-turn interactions and more effective few-shot selection mechanisms. Second, improving confidence estimation in the hybrid architecture could lead to more robust decision-making and better performance trade-offs. Third, we will investigate more sophisticated fusion strategies to better integrate audio and textual modalities. Finally, we will extend the evaluation to additional datasets and languages to assess the robustness and generalization capabilities of ALMs in SER. Overall, this work provides a structured benchmark of ALM-based approaches for emotion classification, highlighting their strengths and limitations and their potential for future research.

Acknowledgement: Not applicable.

Funding Statement: This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU-FEDER/UE)—a way of making Europe. Tomas Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Author Contributions: Jorge Gómez-Navalón: writing—original draft, visualization, validation, software, methodology. Ronghao Pan: writing—review & editing, validation, software, investigation. Tomas Bernal-Beltrán: writing—review & editing, visualization, validation, software, investigation. José Antonio García-Díaz: writing—review & editing, validation, methodology, investigation, formal analysis, data curation, conceptualization. Rafael Valencia-García: writing—review & editing, supervision, project administration, funding acquisition, conceptualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The Spanish MEACorpus dataset used in this study is publicly available in its original publication. Source code, experimental configurations, and fine-tuned models will be released on GitHub and Hugging Face upon acceptance of the paper.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen L, Mao X, Xue Y, Cheng LL. Speech emotion recognition: features and classification models. Digit Signal Process. 2012;22(6):1154–60.
2. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol. 2018;21(1):93–120. doi:10.1007/s10772-018-9491-z.
3. Mao K, Zhang W, Wang DB, Li A, Jiao R, Zhu Y, et al. Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN. IEEE Trans Affect Comput. 2023;14(3):2251–65. doi:10.1109/taffc.2022.3154332.
4. Wang P, Liu A, Sun X. Integrating emotion dynamics in mental health: a trimodal framework combining ecological momentary assessment, physiological measurements, and speech emotion recognition. Interdiscip Med. 2025;3(3):e20240095.
5. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst. 2020;33:12449–60.
6. Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, HuBERT M A. Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3451–60. doi:10.1109/taslp.2021.3122291.
7. Ghosh S, Goel A, Jayakumar K, Koroshinadze L, Anand N, Kong Z, et al. Audio flamingo next: next-generation open audio-language models for speech, sound, and music. Technical report. 2026 [cited 2026 Jan 1]. Available from: https://afnext-umd-nvidia.github.io/.
8. Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, et al. PaLM-E: an embodied multimodal language model. arXiv:2303.03378. 2023.
9. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners. arXiv:2109.01652. 2021.
10. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, et al. A survey on multimodal large language models. Natl Sci Rev. 2024;11(12):nwae403.
11. Yang C-K, Ho NS, H-y L. Towards holistic evaluation of large audio-language models: a comprehensive survey. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025 Nov 4–9; Suzhou, China. p. 10155–81.
12. Pan R, García-Díaz JA, Rodríguez-García MÁ, Valencia-García R. A multimodal speech-text corpus for emotion analysis in Spanish from natural environments. Comput Stand Interfaces. 2023;90(14):103856. doi:10.1016/j.csi.2024.103856 2024.
13. Ekman P. An argument for basic emotions. Cogn Emot. 1992;6(3–4):169–200. doi:10.1080/02699939208411068.
14. Plutchik R. A general psychoevolutionary theory of emotion. In: Theories of emotion. Amsterdam, The Netherlands: Elsevier; 1980. p. 3–33.
15. Singla C, Singh S, Sharma P, Mittal N, Gared F. Emotion recognition for human-computer interaction using high-level descriptors. Sci Rep. 2024;14(1):12122. doi:10.1038/s41598-024-59294-y.
16. Hegde K, Jayalath H. Emotions in the loop: a survey of affective computing for emotional support. arXiv:2505.01542. 2025.
17. George SM, Muhamed Ilyas P. A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. Neurocomputing. 2024;568:127015.
18. Hassan A, Damper RI. Multi-class and hierarchical SVMs for emotion recognition. Proc Interspeech. 2010:2354–7. doi:10.21437/interspeech.2010-644.
19. Subramanian R, Aruchamy P. A survey of human emotion recognition using speech signals: current trends and future perspectives. In: Micro-Electronics and Telecommunication Engineering: Proceedings of 6th ICMETE 2022. Berlin/Heidelberg, Germany: Springer; 2023. p. 509–18.
20. Subramanian R, Aruchamy P. An effective speech emotion recognition model for multi-regional languages using threshold-based feature selection algorithm. Circuits Syst Signal Process. 2024;43(4):2477–506. doi:10.1007/s00034-023-02571-4.

21. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. Speech emotion recognition using deep learning techniques: a review. IEEE Access. 2019;7:117327–45. doi:10.1109/access.2019.2936124.
22. Abbaschian BJ, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models. Sensors. 2021;21(4):1249. doi:10.3390/s21041249.
23. Singh YB, Goel S. A systematic literature review of speech emotion recognition approaches. Neurocomputing. 2022;492(4):245–63. doi:10.1016/j.neucom.2022.04.028.
24. Latif S, Zaidi A, Cuayahuitl H, Shamshad F, Shoukat M, Usama M, et al. Transformers in speech processing: a survey. arXiv:2303.11607. 2023.
25. Zhang H, Huang H, Zhao P, Yu Z. Sparse temporal aware capsule network for robust speech emotion recognition. Eng Appl Artif Intell. 2025;144(10):110060. doi:10.1016/j.engappai.2025.110060.
26. Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE J Sel Top Signal Process. 2022;16(6):1505–18.
27. Chen Z, Liu C, Wang Z, Zhao C, Lin M, Zheng Q. MTLSER: multi-task learning enhanced speech emotion recognition with pre-trained acoustic model. Expert Syst Appl. 2025;273:126855.
28. Pepino L, Riera P, Ferrer L. Emotion recognition from speech using wav2vec 2.0 embeddings. arXiv:2104.03502. 2021.
29. Chakhtouna A, Sekkate S. Unveiling embedded features in Wav2vec2 and HuBERT models for speech emotion recognition. Procedia Comput Sci. 2024;232(5):2560–9. doi:10.1016/j.procs.2024.02.074.
30. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. Int Conf Mach Learn. 2023;202:28492–518.
31. Kaur K, Singh P. Trends in speech emotion recognition: a comprehensive survey. Multimed Tools Appl. 2023;82(19):29307–51. doi:10.1007/s11042-023-14656-y.
32. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech in. In: Proceedings of the Interspeech 2005; 2005 Sep 4–8; Lisbon, Portugal. p. 1517–20.
33. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42(4):335–59.
34. Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13(5):e0196391.
35. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. Meld: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 527–36.
36. Hozjan V, Kacic Z, Moreno A, Bonafonte A, Nogueiras A. Interface databases: design and collection of a multilingual emotional speech database. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02); 2002 May 29–31; Canary Islands, Spain.
37. Garcia-Cuesta E, Salvador AB, Pãez DG. EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database. Multimed Tools Appl. 2024;83(5):13093–112.
38. Kumar MD, Rao MS, Narendra K. Multimodal emotion recognition: a comprehensive survey of datasets, methods, and applications. IEEE Access. 2025;13:201067–97.
39. D'mello SK, Kory J. A review and meta-analysis of multimodal affect detection systems. ACM Comput Surv. 2015;47(3):1–36. doi:10.1145/2682899.
40. Li S, Tang H. Multimodal alignment and fusion: a survey. arXiv:2411.17040. 2024.
41. Wu C, Cai Y, Liu Y, Zhu P, Xue Y, Gong Z, et al. Multimodal emotion recognition in conversations: a survey of methods, trends, challenges and prospects. arXiv:2505.20511. 2025.
42. Baltrusaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43.
43. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion. 2017;37:98–125.

44. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 6558–69.
45. Wang Y, Gu Y, Yin Y, Han Y, Zhang H, Wang S, et al. Multimodal transformer augmented fusion for speech emotion recognition. Front Neurorobot. 2023;17:1181598. doi:10.3389/fnbot.2023.1181598.
46. Zhao Z, Wang Y, Wang Y. Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition. arXiv:2207.04697. 2022.
47. Khan M, Tran P-N, Pham NT, El Saddik A, Othmani A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Sci Rep. 2025;15(1):5473.
48. Zhao J, Li R, Jin Q. Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. p. 2608–18.
49. Al-Azani S, El-Alfy E-SM. A review and critical analysis of multimodal datasets for emotional AI. Artif Intell Rev. 2025;58(10):334. doi:10.1007/s10462-025-11271-1.
50. Yazici A, Kucukyilmaz T, Dokeroglu T, Sharipbay A, Lee M-H, Tyler B. State-of-the-art multimodal emotion recognition: a comprehensive survey and taxonomy. Intell Syst Appl. 2026;30:200642.
51. Bellver J, Martín-Fernández I, Bravo-Pacheco JM, Esteban S, Fernández-Martínez F, D'Haro LF. Multimodal audio-language model for speech emotion recognition. In: Proceedings of the Odyssey 2024: The Speaker and Language Recognition Workshop; 2024 Jun 18–24; Quebec City, QC, Canada. p. 288–95.
52. Zhang D, Li S, Zhang X, Zhan J, Wang P, Zhou Y, et al. SpeechGPT: empowering large language models with intrinsic cross-modal conversational abilities. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore. p. 15757–73.
53. Huang R, Li M, Yang D, Shi J, Chang X, Ye Z, et al. AudioGPT: understanding and generating speech, music, sound, and talking head. Proc AAAI Conf Artif Intell. 2024;38:23802–4.
54. Chu Y, Xu J, Yang Q, Wei H, Wei X, Guo Z, et al. Qwen2-audio technical report. arXiv:2407.10759. 2024.
55. Team G. Gemma 3n. arXiv:2408.00118. 2025.
56. Liu AH, Ehrenberg A, Lo A, Denoix C, Barreau C, Lample G, et al. Voxtral. arXiv:2507.13264. 2025.
57. Team O. MOSS-audio technical report. GitHub repository. 2026 [cited 2026 Jan 1]. Available from: https://github.com/OpenMOSS/MOSS-Audio.
58. Zhao Z, Zhu X, Wang X, Wang S, Geng X, Tian W, et al. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. IEEE Trans Audio Speech Lang Process. 2025;34:415–26. doi:10.1109/taslpro.2025.3648793.
59. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. Iclr. 2022;1(2):3.
60. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901. doi:10.65525/svup.9788199778009.2026.224-230.
61. Hanif A, Agro MT, Qazi MA, Aldarmaki H. PALM: few-shot prompt learning for audio language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12–16; Miami, FL, USA. p. 18527–36.
62. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37. doi:10.52202/068431-1800.
63. Ma Z, Chen Z, Wang Y, Chng E-S, Chen X. Audio-CoT: exploring chain-of-thought reasoning in large audio language model. In: Proceedings of the 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2025 Dec 6–10; Honolulu, HI, USA. p. 1–6.
64. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. doi:10.1145/3571730.
65. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):1–35.

66. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol. 2024;15(3):1–45. doi:10.1145/3641289.
67. Mares A, Diaz-Arango G, Perez-Jacome-Friscione J, Vazquez-Leal H, Hernandez-Martinez L, Huerta-Chua J, et al. Advancing Spanish speech emotion recognition: a comprehensive benchmark of pre-trained models. Appl Sci. 2025;15(8):4340.