EmotionLayer: A Multimodal Architecture for Empathic Voice Assistants based on Speech Emotion Recognition and Large Language Models

Alessio Bernardini[1*]

[1*]Department of Computer Science, Università degli Studi di Milano, Via Celoria, 18, Milan, 20133, MI, Italy.

Corresponding author(s). E-mail(s): alessio.bernardini@studenti.unimi.it;

Abstract

AI-based voice dialogue systems have reached high levels of linguistic competence, yet persist in a critical empathic gap: they process speech as text, discarding the rich paralinguistic information that conveys the user's emotional state. In this work we present **EmotionLayer**, a multimodal architecture that integrates Speech Emotion Recognition (SER) with Large Language Models (LLMs) for empathic voice dialogue in Italian. The system combines emotion classification via Wav2Vec 2.0 fine-tuned on Italian data with continuous prosodic analysis through the PAD model (Pleasure–Arousal–Dominance). We systematically evaluate 18 LLM candidates on their ability to identify the user's emotional state and select the correct response strategy from EmotionLayer's output, using an *LLM-as-a-Judge* methodology on 40 purpose-built gold-standard scenarios. Results show that while proprietary models achieve the highest scores (Gemini 2.5 Flash: 91.7/100; GPT-4.1: 90.3/100), open-source models hosted on Groq offer optimal trade-offs for large-scale production (GPT-OSS-20b: 84.0/100, latency 1.04 s, cost $0.0004 per request). Thanks to the structured emotional context provided by EmotionLayer, the LLM is able to detect complex states such as sarcasm and urgency and generate responses that account for the user's affective state.

Keywords: Speech Emotion Recognition, Large Language Model, empathic dialogue, multimodal architecture, Wav2Vec 2.0, PAD, Italian

1 Introduction

The evolution of Large Language Models (LLMs) has radically transformed the landscape of conversational artificial intelligence, pushing voice assistants toward new frontiers. Until recently, the industry standard relied primarily on modular pipelines composed of disjoint systems for Automatic Speech Recognition (ASR), response processing via LLM, and Text-to-Speech (TTS) synthesis. More recently, however, research has shifted toward end-to-end Speech-to-Speech (S2S) architectures: models that promise reduced latency and greater naturalness by directly handling audio input and output without intermediate text steps.

Despite this technological leap, a fundamental challenge persists: the handling of paralinguistic information. Even the most modern S2S systems, though technically capable of processing audio, tend to focus on semantic content while neglecting the rich information contained in *how* the user speaks. Elements such as tone, rhythm, hesitations, and vocal intensity are rarely encoded and used to enrich context and guide the response. As a result, the system *hears* the audio but does not truly *listen* to the emotional state, losing the context needed to understand deep pragmatic intent — sarcasm, urgency, latent frustration — that transcription alone cannot convey.

This gap becomes a decisive obstacle in numerous critical application domains: from customer service to telemedicine, from digital psychological support to adaptive education, to assistance systems for the elderly and people with disabilities — contexts where understanding emotional context is not an added value, but a fundamental requirement for interaction effectiveness. Responding with the correct strategy depends not only on the technical problem presented, but on the emotional state of the interlocutor: responding to a user exhibiting excitement or anger requires a diametrically opposite communication strategy to one appropriate for a calm user, even given the same semantic request. The absence of *paralinguistic awareness* prevents the assistant from modulating its resolution strategy, risking standardised responses that can exacerbate friction and undermine interaction effectiveness.

1.1 Research Questions

This work investigates the current state of empathic voice assistance in the era of large language models. Specifically, we address the following research questions:

RQ1. How effectively can LLMs detect and respond to complex emotional states when provided with multimodal input (text + prosody)?
RQ2. Which LLM architectures offer optimal trade-offs between emotional intelligence, latency, and operational cost for production deployment?
RQ3. Can current LLMs detect subtle emotional phenomena — sarcasm, urgency, and passive frustration — through prosodic analysis?

1.2 Contributions

To overcome the limitations described, we propose **EmotionLayer**: a specialised module that integrates into the pipeline of a voice assistant by extracting paralinguistic

properties from audio and providing the LLM with a structured representation of the user's emotional state, well beyond simple text transcription.

The main contributions of this work are:

- **SER models for Italian**: fine-tuning of two models derived from `facebook/wav2vec2-large-xlsr-53` — a discrete emotion classifier over 7 classes (trained on Emozionalmente, EMOVO and AI4SER) and a continuous PAD value regressor (trained on AI4SER).
- **EmotionLayer**: integration of the two models into a *production-ready* architecture that fuses discrete classification and continuous prosodic analysis into a structured JSON context, directly usable by any downstream LLM.
- **Systematic benchmark**: evaluation of 18 LLM models on emotional detection capability and strategic selection via *LLM-as-a-Judge* methodology on 40 gold-standard scenarios.
- **Cost–benefit analysis**: guidelines for selecting the optimal LLM model based on the trade-off between emotional quality, latency, and operational cost.

The remainder of this paper is organised as follows. Section 2 introduces the theoretical foundations: the role of paralinguistic information in voice dialogue, and the Ekman and PAD emotional models. Section 3 examines related work in SER, LLMs for empathic dialogue, and multimodal systems. Section 4 describes the EmotionLayer architecture. Section 5 presents the evaluation methodology. Results are reported in Section 6 and discussed in Section 7. Section 8 concludes the work.

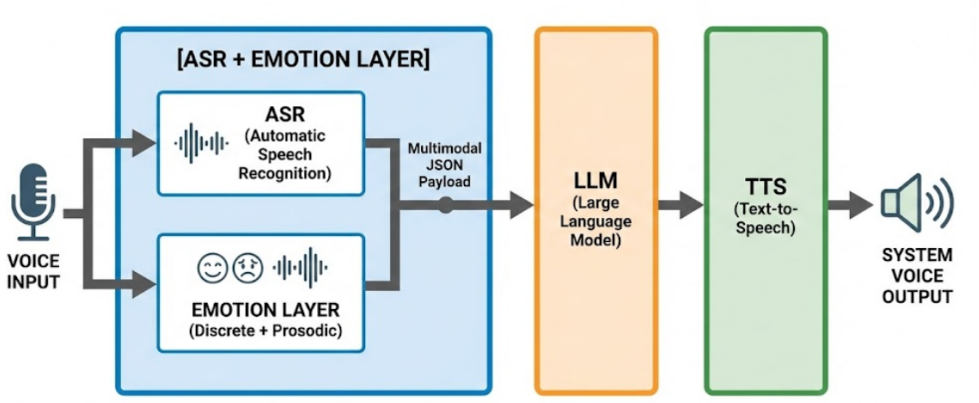

Figure 1: Empathic Vocal Agent pipeline.

2 Background

2.1 Emotional Context in Customer Service

Customer service represents one of the most demanding application domains for conversational AI systems. Interaction quality depends not solely on the correctness of the response provided, but to a significant degree on the system's ability to adapt its communication strategy to the emotional state of the interlocutor. Established key performance indicators in the sector — such as the Customer Satisfaction Score (CSAT), churn rate, and First Call Resolution (FCR) — are strongly influenced by the customer's perception of empathy, independently of the technical outcome of the request.

A customer expressing frustration or urgency requires a structurally different response from one posing the same question in a neutral tone: not only in content, but in rhythm, register, and the de-escalation strategy adopted. Current systems, processing exclusively transcribed text, treat these two situations as equivalent, producing standardised responses that risk exacerbating tension rather than reducing it. The lack of *paralinguistic awareness* therefore constitutes a structural limitation, not resolvable by merely increasing the capacity of language models.

2.2 Paralinguistic Information as an Emotional Signal

Human communication transmits meaning through two distinct and complementary channels: the verbal channel, which conveys semantic content, and the paralinguistic channel, which encodes information relating to emotional state, communicative intent, and relational context. The main paralinguistic features relevant to emotional state recognition include:

- **Pitch** (fundamental frequency F_0): correlated with emotional activation; high and variable values are associated with excitement, anger or anxiety, while low and stable values indicate calm or sadness.
- **Energy and intensity**: the amplitude of the signal reflects the level of emotional engagement; high-energy speech is typically associated with anger or enthusiasm.
- **Rhythm and speech rate**: urgency tends to accelerate speech, while sadness or uncertainty slow it down, introducing pauses and hesitations.
- **Voice quality**: parameters such as jitter, shimmer, and the harmonic-to-noise ratio (HNR) characterise emotional states such as tension or crying.

When an ASR system converts the audio signal to text, this information is irrecoverably lost. The transcription *"Sure, I'll wait"* does not distinguish between a serene affirmation and a sarcastic response laden with frustration, despite being acoustically very different. It is precisely this informational loss that EmotionLayer aims to bridge, operating directly on the audio signal before transcription.

2.3 Ekman's Categorical Model

The most influential model in the psychology of emotions is that proposed by Ekman [1], which identifies six universal basic emotions — joy, sadness, anger, fear, surprise and disgust — cross-culturally recognisable through characteristic facial expressions. This categorical model strongly influenced the first generation of Speech Emotion Recognition systems, which tended to formulate the problem as a multi-class classification over these discrete categories.

Despite its theoretical robustness and wide adoption, Ekman's model presents significant limitations for real-world customer service applications. First, human emotions rarely manifest in pure form: a single utterance may contain components of multiple emotions simultaneously. Second, and more critically, complex and pragmatically relevant emotional states — such as **sarcasm**, **uncertainty**, **passive frustration** or **urgency** — do not correspond to any of the six basic categories, making them difficult to represent in this discrete space. These nuances, however, are precisely those that a voice assistant must be able to recognise in order to respond appropriately.

2.4 The PAD Model: Continuous Dimensional Representation

To overcome the limitations of the categorical model, cognitive psychology has developed dimensional representations of the emotional space. The PAD model (*Pleasure–Arousal–Dominance*), proposed by Russell and Mehrabian [2], describes any affective state as a point in a continuous three-dimensional space, defined by three independent axes:

- **Valence** (Pleasure, V): the positive–negative axis of emotional experience, ranging from unpleasant states ($V = -1$) to pleasant ones ($V = +1$).
- **Arousal** (Activation, A): the level of physiological excitement, from calm and passive states ($A = -1$) to activated and alert ones ($A = +1$).
- **Dominance** (Control, D): the perceived sense of control over the situation, from submissive states ($D = -1$) to controlling ones ($D = +1$).

The power of the PAD model lies in its ability to represent complex emotional states that elude discrete categorisation. Sarcasm, for example, can be approximated by a combination of negative Valence, moderate Arousal and high Dominance — the speaker controls the situation but experiences a negative emotion not directly expressed. Passive frustration is characterised by negative Valence, low Arousal and low Dominance, clearly distinguishable from explicit anger (negative Valence, high Arousal). This representational granularity makes the PAD model particularly suited to guiding adaptive response generation in an LLM, providing a structured and continuous emotional context in place of a simple categorical label.

2.5 Adaptive Communicative Strategies in Emotional Dialogue

Selecting the appropriate communication strategy based on the interlocutor's emotional state is a central problem in communication psychology, pragmatic linguistics, and service science. The literature converges on the idea that an effective response

is not merely content-correct, but adapts register, tone, and structure to the affective context of the speaker. The six strategies adopted in EmotionLayer each have a consolidated theoretical foundation, summarised here.

Active De-escalation (Anger, Frustration). In verbal conflict contexts typical of customer service, de-escalation represents the first-line intervention recommended by the literature. Techniques such as active listening, explicit validation of the customer's frustration, and adoption of a slowed and calm tone significantly reduce perceived tension and increase the probability of first-contact resolution [3]. The primary objective is not to solve the technical problem, but to lower the speaker's Arousal before proceeding with the response content.

Empathic Validation (Uncertainty, Insecurity). For states characterised by low dominance and negative valence, the optimal strategy is empathic validation, a communicative pattern derived from Motivational Interviewing (MI) by Miller and Rollnick [4]. MI is grounded in a collaborative and non-judgmental approach that explicitly acknowledges the speaker's concerns before proposing solutions, producing significantly better outcomes than the directive approach and increasing the perception of being understood.

Positive Mirroring (Enthusiasm, Satisfaction). For positive emotional states, the optimal strategy is emotional mirroring: adapting tone, rhythm and register of the response to the speaker's affective state. Chartrand and Bargh [5] demonstrated that behavioural mimicry — the so-called *chameleon effect* — increases interaction fluidity and the perception of mutual likeability. In a vocal context, this translates into an increased speech rate and adoption of an enthusiastic tone that strengthens *rapport* and amplifies positive emotional resonance.

Assertive Neutrality (Sarcasm). Sarcasm is an *off-record* speech act in the taxonomy of Brown and Levinson [6]: it threatens the interlocutor's positive *face* by conveying an intended meaning opposite to the literal one. The optimal response consists neither in ignoring the sarcastic component nor in reacting to the provocation, but in adopting a professional, direct and non-reactive register that returns the interaction to the cooperative plane without amplifying relational tension.

Rapid Resolution (Urgency). States of urgency are characterised by high arousal and low dominance: the speaker perceives time pressure or a critical situation over which they have little control. The crisis communication literature indicates that the effective response must be rapid, structured and action-oriented [7]. An accelerated and direct delivery signals competence and readiness, reducing the perception of waiting and restoring the speaker's sense of control over the situation.

Direct Efficiency (Neutral). For neutral emotional states, the optimal strategy is task-oriented communication: concise, informative, and free of emotional redundancy. This choice is consistent with Grice's maxim of quantity [8], according to which an effective communicative contribution must contain exactly the information necessary for the conversational purpose, without adding superfluous affective charge that would appear artificial in the absence of a relevant emotional state to mirror.

3 Related Work

3.1 Speech Emotion Recognition

Speech Emotion Recognition (SER) has undergone rapid evolution: from early architectures based on handcrafted features (MFCC, F_0, energy) fed to SVM or HMM classifiers, to CNN-LSTM networks learning representations directly from spectrograms, to *self-supervised* Transformer-based models. Wav2Vec 2.0 [9], pre-trained on thousands of hours of unlabelled speech, represents the state of the art: its supervised fine-tuning surpasses previous models on the IEMOCAP [10] benchmark with a 7.4 percentage point gain in unweighted accuracy. Successor models such as WavLM [11] and HuBERT [12] have further consolidated this paradigm across multiple multilingual benchmarks.

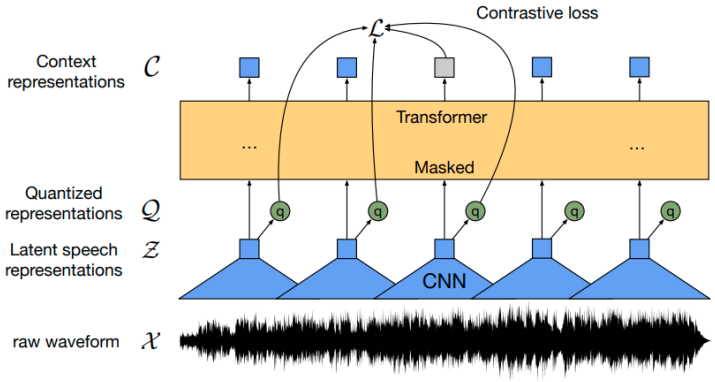

Figure 2: Wav2Vec 2.0 architecture.

3.2 Datasets for Emotional Recognition in Italian

The availability of resources for SER in Italian remains limited compared to major languages: most reference benchmarks (IEMOCAP [10], RAVDESS [13], CREMA-D [14]) are in English, and a recent survey has catalogued only 66 Italian speech datasets, a minority of which are dedicated to emotional analysis [15]. The three main corpora used in this work are as follows.

1. **EMOVO** [16] (2014): the first Italian emotional corpus, with 588 samples recorded by 6 professional actors across 7 emotional categories and 80% perceptual accuracy.
2. **Emozionalmente** [17] (2021): the largest available corpus, with 6,902 samples from 431 non-professional speakers collected via crowdsourcing. A fine-tuned Wav2Vec 2.0 model achieves 82.45% accuracy. Available on Zenodo under an open-source licence.
3. **AI4SER** [18]: the only Italian dataset with native PAD annotations (*Valence*, *Arousal*, *Dominance*) per sample, available on Hugging Face under CC-BY 4.0 licence (DOI: `10.57967/hf/6703`). Contains 3,500 audio samples.

The prevalence of *acted* corpora and the absence of datasets covering complex emotional states beyond Ekman's categories motivate the choice of using all three corpora in a complementary manner for EmotionLayer fine-tuning.

3.3 LLMs for Empathic Dialogue

The application of Large Language Models to empathic dialogue has expanded rapidly since 2023, driven by the availability of high-capacity models such as GPT-4 [19] and growing demand for conversational agents in emotionally intensive domains — mental health, psychological support, and customer service. LLMs demonstrate notable capability in generating responses perceived as empathic by users, thanks to their exposure during pre-training to a wide variety of emotionally connoted conversational texts.

Over the years, the first benchmarks emerged for systematically evaluating the emotional intelligence of LLMs. The main ones are:

1. **EQ-Bench** [20] (Paech, 2023): evaluates emotional comprehension by asking LLMs to predict the intensity of emotional states of characters in complex dialogues. The third version, **EQ-Bench 3** [21], adopts multi-turn role-play scenarios judged via LLM-as-a-Judge, measuring empathy, social intuition, and interpersonal dexterity.
2. **EmotionQueen** [22] (Chen et al., 2024): a framework focused on empathic response generation, articulated across four tasks — recognition of key events, mixed events, implicit emotions, and intentions — with metrics evaluating recognition capability and response quality separately.

A parallel strand has explored prompt engineering as a tool for inducing empathic behaviours in LLMs. Arjmand et al. [23] introduced the concept of *empathic grounding* for conversational agents, demonstrating that a multimodal model integrating speech and facial expressions via GPT-3.5 significantly increases the perception of empathy, understanding, and trust on the part of users compared to systems generating only non-affective signals. More recent contributions, such as the *Empathic Prompting* framework [24], extended this approach by integrating valence and arousal, derived via facial expression recognition, into the language model's prompt. This last architecture is close to EmotionLayer's approach, which distinguishes itself through the use of the audio signal in place of facial expressions.

Regarding the automatic evaluation of empathy, the *LLM-as-a-Judge* framework — adopted in the present work — was validated by Zheng et al. [25] on MT-Bench and Chatbot Arena, demonstrating high agreement with human evaluations for open-ended dialogue tasks.

3.4 Multimodal Systems for Emotion Recognition

Multimodal Emotion Recognition (MER) is growing rapidly, with nearly 80% of contributions having appeared after 2019 [26]. The dominant trend is the shift from early/late fusion toward cross-modal Transformer-based architectures, evaluated primarily on IEMOCAP [10] and MELD [27]. Wang et al. [28] propose a systematic

benchmark of fine-tuning Wav2Vec 2.0 and HuBERT for emotional recognition, achieving state-of-the-art results. Siriwardhana et al. [29] demonstrate that joint audio-text fine-tuning consistently outperforms separate fusion across multiple datasets.

EmotionLayer instead adopts a modular integration: the two signals are processed separately and combined into a JSON context provided to a generalist LLM. This approach prioritises flexibility, interpretability, and replaceability of the LLM component — fundamental requirements in industrial environments where models evolve rapidly.

4 System Architecture

EmotionLayer is an emotional analysis module designed to operate in parallel with automatic speech transcription within a voice customer service pipeline. Given an audio file as input, the system simultaneously performs three operations: discrete classification of the dominant emotion, estimation of PAD dimensional values, and text transcription. The three streams are then fused into a single structured JSON object, which constitutes the enriched emotional context provided to the downstream LLM for response generation.

Figure 3 illustrates the overall system pipeline.

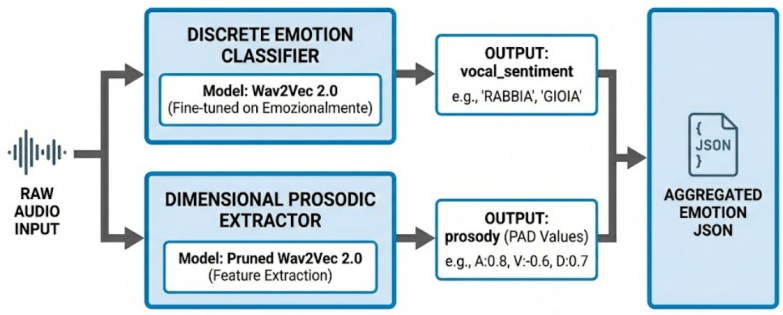

Figure 3: EmotionLayer architecture pipeline.

Both neural models of EmotionLayer share the same base model: `facebook/wav2vec2-large-xlsr-53` [9], a multilingual Transformer pre-trained on 53 languages with 24 Transformer layers, 16 attention heads, hidden size of 1024, and 7 convolutional layers for low-level acoustic feature extraction. The choice of a large-scale pre-trained multilingual model guarantees robust and generalisable audio signal representations, reducing dependence on the quantity of fine-tuning data available for Italian.

4.1 Discrete Emotion Classifier

The first EmotionLayer component is a classifier that assigns a discrete emotional label to the input audio segment from the seven defined classes: *anger*, *disgust*, *fear*, *joy*, *neutrality*, *sadness* and *surprise* — corresponding to Ekman's six emotions [1]

plus the neutral class. The model adopts mean pooling (`pooling_mode = mean`) over the contextualised representations of the Wav2Vec 2.0 encoder, followed by a linear classification head projecting into a 7-dimensional space.

4.1.1 Training Data

The classifier was trained on the union of the three Italian datasets described in Section 3.2: Emozionalmente, EMOVO and AI4SER. The combined dataset comprises 30,878 training samples and 3,860 validation samples. This aggregation strategy aims to maximise stylistic and demographic coverage, compensating for the dimensional limitations of individual corpora.

4.1.2 Training Configuration

Fine-tuning was performed with the following configuration:

Table 1: Training hyperparameters for the discrete classifier.

Parameter	**Value**
Base model	`facebook/wav2vec2-large-xlsr-53`
Batch size (train)	32
Batch size (eval)	32
Gradient accumulation	4 steps (effective batch: 128)
Learning rate	1×10^{-4}
Epochs	20 (best checkpoint at epoch 10)
Numerical precision	FP16
Feature extractor	Frozen
Saving strategy	Best model on eval loss

4.1.3 Results

The model reached at the best checkpoint (epoch 10, step 2000) an evaluation accuracy of **66.81%** on 3,860 test samples, with an evaluation loss of 0.988. Training loss settled at 1.009, over a total of 30,878 samples processed in approximately 4 hours and 25 minutes (19.41 samples/s). The modest gap between train loss and eval loss indicates absence of marked overfitting. The accuracy reflects the intrinsic difficulty of the task on heterogeneous multi-corpus data, where stylistic distributions of individual datasets may introduce labelling noise.

4.2 PAD Prosodic Regressor

The second component continuously estimates the three PAD model values — Valence (V), Arousal (A) and Dominance (D) — directly from the audio signal. Unlike the discrete classifier, this model addresses a multi-target regression task: the output is a vector $(V, A, D) \in \mathbb{R}^3$ representing the utterance's position in the three-dimensional emotional space.

4.2.1 Training Data

The regressor was trained exclusively on the AI4SER dataset, the only one among the available Italian corpora to provide PAD dimensional annotations per sample. The dataset was split into 2,800 training samples and 700 validation samples.

4.2.2 Training Configuration

Table 2: Training hyperparameters for the PAD regressor.

Parameter	Value
Base model	`facebook/wav2vec2-large-xlsr-53`
Batch size (train)	32
Batch size (eval)	16
Gradient accumulation	4 steps (effective batch: 128)
Learning rate	1×10^{-4}
Epochs	30
Numerical precision	BF16
Feature extractor	Frozen
Output	Multi-target regression (V, A, D)

4.2.3 Results

Regressor performance is evaluated via the Concordance Correlation Coefficient (CCC), the standard metric for emotional regression tasks as it combines correlation with scale agreement. Results obtained at epoch 30 are reported in Table 3.

Table 3: PAD regressor performance on AI4SER (validation set).

Dimension	CCC	MSE
Arousal	0.615	
Dominance	0.661	0.026
Valence	0.522	
Average	**0.599**	**0.026**

The mean CCC of 0.599 indicates moderate-good agreement, in line with the state of the art for models trained on Italian corpora of limited size. Dominance achieves the highest value (0.661), while Valence presents the most uncertain estimate (0.522) — a result consistent with the literature, which systematically reports valence as the most difficult dimension to predict from the acoustic signal alone.

4.3 Fusion and Integration with the Large Language Model

The two models operate in parallel on the same audio segment, so as to minimise the overall system latency. Their outputs, together with the text transcription produced by the ASR module, are fused into a structured JSON object with the following schema:

Listing 1: EmotionLayer JSON output schema.

```
{
  "vocal_sentiment": [
    { "Emotion": "anger", "Score": "75%" },
    { "Emotion": "sad",   "Score": "10%" }
  ],
  "prosody": {
    "arousal":   0.95,
    "dominance": 0.90,
    "valence":   0.05
  },
  "transcription": "..."
}
```

The `vocal_sentiment` field reports the discrete emotions with their confidence scores in descending order, allowing the LLM to distinguish the dominant emotion from secondary emotions present. The `prosody` field provides the continuous PAD values, enabling characterisation of nuances not capturable from the categorical label alone: in the example shown, Arousal= 0.95 and Valence= 0.05 signal a state of high negative activation, while Dominance= 0.90 indicates that the speaker perceives themselves to be in control of the situation — a signature typical of assertive anger rather than passive frustration.

This JSON is inserted into the LLM's system prompt as additional context, instructing the model to adapt its response strategy based on the detected emotional state. The separation between discrete classification and dimensional estimation allows the LLM to reason both qualitatively ("the user is angry") and quantitatively ("with high arousal and low valence"), maximising the richness of emotional context available for generation.

5 Evaluation Methodology

The objective of the evaluation is to determine the extent to which leading available LLMs can leverage EmotionLayer's structured output to select the emotionally appropriate response strategy and generate contextually adequate responses. The adopted framework consists of three elements: a set of manually constructed gold-standard scenarios, an automatic evaluation mechanism based on *LLM-as-a-Judge*, and a weighted scoring system across six metrics.

5.1 Emotion and Strategy Taxonomy

The system distinguishes nine detectable emotional states. The categories do not correspond solely to Ekman's basic emotions, but include states derived through semantic composition: each is defined by its origin in primary emotions, by quantitative thresholds on PAD values, and, in the case of sarcasm, by a discrepancy between textual content and acoustic signal. Table 4 provides the complete formal definition.

Table 4: Emotion taxonomy: semantic composition and associated PAD thresholds.

Emotion	Composition	Valence	Arousal	Dominance
Neutral	Neutral	≈ 0.5	≈ 0.5	—
Enthusiasm	Joy + Surprise	> 0.7	> 0.7	—
Satisfaction	Joy	> 0.6	< 0.5	—
Uncertainty	Fear + Surprise	< 0.5	≈ 0.5	< 0.4
Anger	Anger	< 0.3	> 0.8	> 0.7
Frustration	Anger + Sadness	< 0.4	> 0.6	< 0.5
Sarcasm	Disgust + Anger	Text/audio discrepancy		
Urgency	Fear + Anger	< 0.4	> 0.8	< 0.5
Insecurity	Fear + Sadness	< 0.4	< 0.4	< 0.3

Each emotional state corresponds to a predefined response strategy that specifies the communication register to adopt and the parameters to provide to the TTS module for modulating speed, tone and pitch of the synthesised voice. The six available strategies are described in Table 5.

Table 5: Response strategies and associated TTS configuration.

Strategy	Target emotions	Speed	Tone	Pitch
Direct Efficiency	Neutral	1.00	informative	balanced
Active De-escalation	Anger, Frustration	0.80	apologetic	low
Empathic Validation	Insecurity, Uncertainty	0.85	empathetic	low
Positive Mirroring	Enthusiasm, Satisfaction	1.10	cheerful	high
Assertive Neutrality	Sarcasm	1.00	professional	low
Rapid Resolution	Urgency	1.25	professional	default

Each model is required to produce a structured JSON output containing the synthetic reasoning, selected strategy, detected emotion, TTS configuration, and response text:

Listing 2: Expected JSON output schema for LLM models.

```
{
  "reasoning": "...",
  "selected_strategy": "CHOSEN_STRATEGY",
  "emotion_detected": "DETECTED_EMOTION",
  "tts_config": {
    "speed": 0.8,
    "tone": "apologetic",
    "pitch": "low"
  },
  "response_text": "The response text."
}
```

5.2 Gold-Standard Scenarios

Forty gold-standard test cases were manually constructed, each simulating a realistic voice customer service interaction. Each scenario consists of a JSON object in the form produced by EmotionLayer — containing the fields `vocal_sentiment`, `prosody` and `transcription` — and a reference annotation specifying the expected emotion and the correct strategy to adopt.

Manual construction of scenarios was preferred over using existing datasets to ensure coverage of the entire emotional space defined by the system, including complex emotions that available corpora do not systematically cover. Each scenario was designed so that the correct emotion cannot be inferred from the text transcription alone, but requires integration with prosodic signals: this ensures that the evaluation actually measures the LLM's ability to reason in a multimodal manner.

The following excerpt illustrates an example scenario with a high-intensity anger emotion:

Listing 3: Example gold-standard scenario (anger).

```
{
  "id": "ecom_ang_01",
  "target_strategy": "ACTIVE DE-ESCALATION",
  "emotion_detected": "ANGER",
  "input": {
    "vocal_sentiment": [
      { "Emotion": "anger", "Score": "96%" },
      { "Emotion": "disgust", "Score": "3%" }
    ],
    "prosody": {
      "arousal": 0.95, "dominance": 0.90, "valence": 0.05
    },
    "transcription": "The courier wrote that I wasn't home
```

```
      but I've been here since this morning! It's false!"
  }
}
```

5.3 LLM-as-a-Judge Framework

Evaluation of responses produced by the 18 candidate models is entrusted to an automatic judge implemented with **Claude Opus 4.6** (`claude-opus-4-6`), chosen as the model with the best performance on reasoning benchmarks available at the time of evaluation. Using an LLM as judge [25] enables scaling the evaluation across the full set of 40 scenarios $\times$ 18 models $=$ 720 responses, while preserving qualitative assessment of dimensions not capturable by traditional automatic metrics.

The judge assigns six metrics to each response, organised in two groups:

- **Binary metrics** (combined weight 40%):
  - *Emotion Match* (0/1): correct identification of emotion from the multimodal context.
  - *Strategy Match* (0/1): correct selection of the psychological strategy.
- **Continuous metrics** (combined weight 60%, scale 1–5):
  - *Relevance*: relevance of the response to the specific problem presented.
  - *TTS Alignment*: consistency of TTS parameters with the detected emotional state.
  - *Voice Suitability*: suitability of the text for speech synthesis (conciseness, natural phrasing).
  - *Empathic Response*: quality of expressed empathy, evaluated according to the PERM framework.

The final normalised score on a 0–100 scale is calculated as:

$$S = (0.20 \cdot E_m + 0.20 \cdot S_m + 0.15 \cdot R + 0.10 \cdot T + 0.10 \cdot V + 0.25 \cdot P) \times 100 \qquad (1)$$

where E_m and S_m are the binary metrics normalised to $[0, 1]$, and R, T, V, P are the continuous metrics normalised to $[0, 1]$ from the 1–5 scale.

5.4 Candidate Models and Experimental Configuration

Eighteen LLM models from five providers were evaluated, selected on the basis of high performance on the main available emotional benchmarks (EQ-Bench 3 [21] and EmotionQueen [22]) and to cover a broad spectrum of sizes, architectures, and cost ranges. Table 6 provides the complete list.

All models were queried with the same configuration: temperature $= 0$ to ensure reproducibility of responses, no stochastic sampling, and a uniform system prompt providing the emotion taxonomy, strategy definitions, and the expected JSON output schema. Response latency and cost per request were recorded for each model to construct the cost–benefit analysis presented in Section 6.

Table 6: Candidate LLM models for evaluation.

Model	Provider	Infrastructure
gpt-5-mini-2025-08-07	OpenAI	OpenAI API
gpt-5.2-2025-12-11	OpenAI	OpenAI API
gpt-4.1-2025-04-14	OpenAI	OpenAI API
gpt-4.1-mini	OpenAI	OpenAI API
Qwen2.5-7B-Instruct	HuggingFace / Together	Together API
Qwen2.5-72B-Instruct	HuggingFace / Novita	Novita API
Mistral-7B-Instruct-v0.2	HuggingFace / Featherless	Featherless API
openai/gpt-oss-120b	OpenAI OSS	Groq
openai/gpt-oss-20b	OpenAI OSS	Groq
llama-3.1-8b-instant	Meta	Groq
llama-3.3-70b-versatile	Meta	Groq
qwen/qwen3-32b	Alibaba	Groq
claude-sonnet-4-5	Anthropic	Anthropic API
claude-haiku-4-5	Anthropic	Anthropic API
claude-opus-4-5	Anthropic	Anthropic API
gemini-2.5-flash-preview	Google	Google API
gemini-2.5-pro	Google	Google API
gemini-2.5-flash-lite	Google	Google API

6 Results

6.1 Overall Ranking

Table 7 reports the overall scores obtained by the 18 evaluated models, sorted by descending score, together with average latency per request and estimated cost per single inference. Table 8 details the six individual metrics for each model.

Results show a clear stratification into three tiers. The top tier, comprising three models with scores above 90 — Gemini 2.5 Flash Preview (91.7), Gemini 2.5 Pro (90.5) and GPT-4.1 (90.3) — is characterised by emotion identification and strategy selection rates of 0.90–0.92, indicating robust understanding of the multimodal context produced by EmotionLayer. The middle tier, between 84 and 89.5 points, gathers heterogeneous models by provider and size — from Claude Sonnet 4.5 to GPT-OSS-20b on Groq — sharing a good balance between emotional accuracy and response quality. The bottom tier, below 75 points, comprises the smaller models (Qwen 7B, Llama 3.1 8B, Mistral 7B) and the Mistral-7B-Instruct-v0.2 model which closes the ranking at 64.85 points.

An unexpected result concerns OpenAI's two GPT-5 models: despite being the more recent generation, **gpt-5.2** (78.62) and **gpt-5-mini** (80.30) rank in the lower-middle tier, well below GPT-4.1 (90.3). Analysis of the detailed metrics reveals the cause: both models exhibit anomalously low *Voice Suitability* scores (1.87 and 2.12 respectively), suggesting that GPT-5 models tend to produce verbose or stylistically unsuitable texts for speech synthesis. This result highlights that a newer generation does not necessarily imply better performance on specialised tasks such as empathic voice assistance.

Table 7: Performance comparison: Quality, Latency, Cost (sorted by descending score).

Provider	Model	Score	Latency (s)	Cost ($)
Gemini	gemini-2.5-flash-preview	**91.70**	4.328	0.001000
Gemini	gemini-2.5-pro	**90.50**	10.617	0.003944
OpenAI	gpt-4.1-2025-04-14	**90.30**	3.769	0.004647
Anthropic	claude-sonnet-4-5	89.46	8.475	0.010305
Anthropic	claude-opus-4-5	87.64	8.270	0.016995
Groq	openai/gpt-oss-120b	87.10	1.594	0.000633
Groq	qwen/qwen3-32b	86.20	1.623	0.000819
HuggingFace	Qwen2.5-72B-Instruct	85.60	8.927	0.000647
Gemini	gemini-2.5-flash-lite	85.40	1.494	0.000235
Groq	openai/gpt-oss-20b	84.00	1.041	0.000364
OpenAI	gpt-4.1-mini	82.75	3.362	0.000909
Groq	llama-3.3-70b-versatile	82.60	0.758	0.001051
OpenAI	gpt-5-mini-2025-08-07	80.30	19.667	0.003060
Anthropic	claude-haiku-4-5	80.06	3.909	0.003355
OpenAI	gpt-5.2-2025-12-11	78.62	5.769	0.007106
HuggingFace	Qwen2.5-7B-Instruct	73.40	2.670	0.000509
Groq	llama-3.1-8b-instant	71.20	0.511	0.000217
HuggingFace	Mistral-7B-Instruct-v0.2	64.85	12.180	0.000476

6.2 Quality, Latency and Cost Trade-off Analysis

The data in Table 7 reveal markedly different trade-off dynamics across providers. Gemini 2.5 Pro, while achieving second place overall for quality (90.5), presents the highest latency of the benchmark (10.617 s) and a unit cost of $0.003944, making it less suitable for high-frequency production environments. At the opposite end of the spectrum, **gpt-5-mini** registers the highest absolute latency (19.667 s) while placing in the lower-middle quality tier (80.30), making it the least convenient choice in the entire benchmark on both dimensions.

The optimal trade-off for production deployment emerges clearly: **GPT-4.1** (90.3 points, 3.769 s, $0.004647) offers top-tier quality with contained latency and maximum TTS Alignment (4.85). For applications with stringent cost constraints, the landscape divides into two complementary alternatives: **Gemini 2.5 Flash Lite** (85.4 points, 1.494 s, $0.000235) for the best quality-to-cost ratio among proprietary models, and **openai/gpt-oss-20b** on Groq (84.0 points, 1.041 s, $0.000364) as an open-source alternative with the lowest latency among mid-range models.

A particularly relevant result for open-source deployment is **openai/gpt-oss-120b** on Groq: with 87.1 points, a latency of 1.594 s and a cost of $0.000633, it ranks sixth overall, surpassing proprietary top-tier models such as Claude Opus 4.5 (87.64, $0.016995) at 27 times lower cost. The Emotion–Strategy pair of gpt-oss-120b (0.90/0.90) is identical to that of the leading models, making it the most efficient choice in the entire benchmark for environments with simultaneous cost and latency constraints.

Table 8: Detailed evaluation metrics for each model.

Model	Score	Emotion Match (0/1)	Strategy Match (0/1)	Relevance (1–5)	TTS Align. (1–5)	Voice Suit. (1–5)	Empathic Resp. (1–5)
gemini-2.5-flash-preview	91.70	0.90	0.92	4.92	4.75	4.70	4.30
gemini-2.5-pro	90.50	0.90	0.90	4.90	4.68	4.82	4.18
gpt-4.1-2025-04-14	90.30	0.90	0.92	4.92	4.85	4.70	3.90
claude-sonnet-4-5	89.46	0.86	0.89	4.97	4.84	4.24	4.32
claude-opus-4-5	87.64	0.82	0.85	4.94	4.64	4.67	4.33
openai/gpt-oss-120b	87.10	0.90	0.90	4.88	4.78	3.90	3.75
qwen/qwen3-32b	86.20	0.80	0.85	4.80	4.78	4.68	4.05
Qwen2.5-72B	85.60	0.80	0.82	4.75	4.78	4.55	4.18
gemini-2.5-flash-lite	85.40	0.75	0.85	4.78	4.82	4.60	4.12
openai/gpt-oss-20b	84.00	0.82	0.88	4.65	4.62	4.20	3.58
gpt-4.1-mini	82.75	0.75	0.80	4.78	4.60	4.60	3.95
llama-3.3-70b	82.60	0.75	0.80	4.68	4.80	4.62	3.85
gpt-5-mini	80.30	0.85	0.88	4.72	4.58	2.12	3.55
claude-haiku-4-5	80.06	0.71	0.71	4.88	4.53	3.97	4.50
gpt-5.2	78.62	0.82	0.85	4.79	4.51	1.87	3.62
Qwen2.5-7B	73.40	0.70	0.68	3.98	4.20	4.45	3.48
llama-3.1-8b	71.20	0.65	0.62	3.95	4.40	4.68	3.35
Mistral-7B	64.85	0.57	0.65	3.40	3.98	4.05	2.75

La Figura 4 visualises the quality–latency trade-off for all 18 evaluated models, highlighting the optimal zone in the top-left corner (high quality, low latency).

6.3 Analysis by Metric Dimension

Analysis of individual metrics reveals cross-provider patterns. *Relevance* is the most uniformly high dimension: all models scoring above 80 obtain values between 4.65 and 4.97, confirming that comprehension of the semantic content of the request is not the differentiating factor. The main discriminant is the Emotion–Strategy pair, ranging from 0.57/0.65 (Mistral-7B) to 0.90/0.92 (leading models) — a 58% gap in emotional identification.

The *Voice Suitability* metric reveals the most interesting and unexpected pattern. OpenAI's GPT-5 models obtain scores of 1.87 and 2.12 — significantly lower than even smaller models such as Llama 3.1 8B (4.68) and Mistral-7B (4.05). This data suggests that optimisation toward long and articulated responses, typical of newer-generation models, penalises the naturalness of synthesised speech, where short sentences and conversational rhythm are preferable.

Claude Haiku 4.5 confirms the anomalous pattern already emerged in previous results: despite being the worst-performing Anthropic model on Emotion (0.71) and Strategy (0.71), it obtains the highest Empathy score of the entire benchmark (4.50), surpassing Claude Opus 4.5 (4.33) and Claude Sonnet 4.5 (4.32). This profile — high

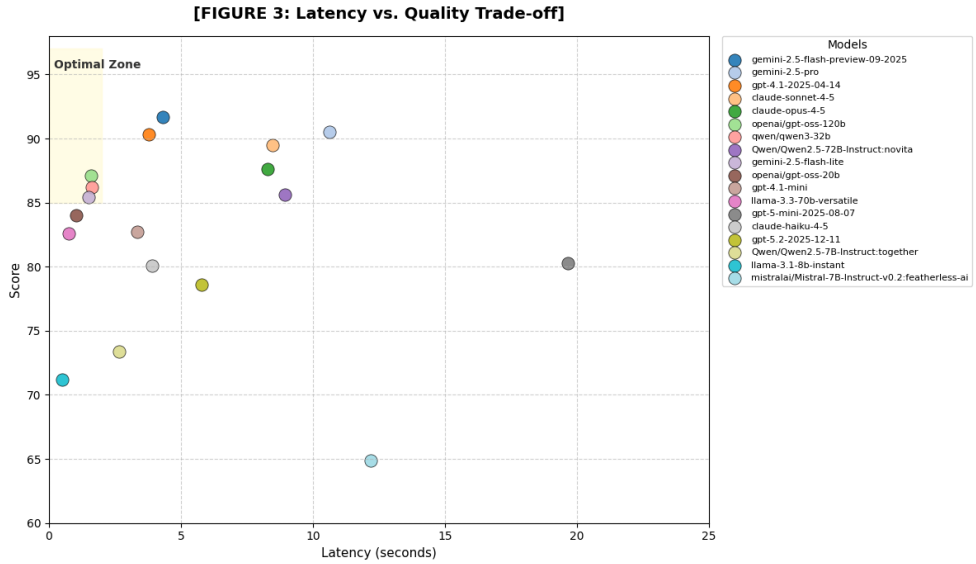

Figure 4: Quality vs. Latency trade-off for the 18 evaluated LLM models. The optimal zone (top-left, highlighted in yellow) identifies models with high score and low latency. Gemini 2.5 Flash Preview positions in the optimal quadrant with the highest score (91.7) and a latency of 4.3 s; GPT-OSS-20b on Groq offers the best open-source trade-off (84.0 points, 1.04 s). GPT-5 models show high latency anomalies against scores below expectations.

empathy, low diagnostic accuracy — indicates a model that produces warm and reassuring responses independently of the correctness of the emotional diagnosis, a risky behaviour in urgency or escalation scenarios.

Claude Sonnet 4.5 emerges instead as the model with the highest *Relevance* of the entire benchmark (4.97), confirming deep understanding of the application context, balanced by good empathy (4.32) and high TTS Alignment (4.84).

7 Discussion

7.1 Answers to Research Questions

RQ1 — LLM effectiveness in understanding complex emotional states with multimodal input. Results demonstrate that top-tier models are able to effectively leverage the structured emotional context provided by EmotionLayer. The three models with scores above 90 achieve emotion identification rates of 90% and strategy selection rates of 90–92%, on scenarios that include complex states such as sarcasm, urgency, and passive frustration — states that, as discussed in Section 2.3, are not representable in Ekman's categories [1] and require integration of prosodic signals to be distinguished. This result affirmatively answers RQ1: current top-tier LLMs *are* capable of reasoning about complex emotional states, provided they receive a structured and semantically rich multimodal context such as that produced by EmotionLayer.

RQ2 — LLM architectures with optimal trade-offs for production. The quality–latency–cost trade-off analysis produces a differentiated answer depending on the deployment context. For quality-priority environments, **GPT-4.1** (90.3 points, 3.769 s, $0.004647) represents the optimal trade-off among proprietary models, combining top-tier quality with acceptable latency for real-time voice interactions. For deployments with stringent cost constraints, **openai/gpt-oss-120b** on Groq (87.1 points, 1.594 s, $0.000633) emerges as the most efficient choice in the entire benchmark, with the same emotional recognition performance as the leading models at 7 times lower cost than GPT-4.1. The Groq ecosystem is confirmed as the reference infrastructure for open-source deployment, with latencies systematically below 2 s even for large models.

RQ3 — Detection of subtle emotional phenomena via prosodic analysis. The system demonstrates the ability to detect non-trivial emotional states. In particular, recognition of **sarcasm** — the most difficult nuance to identify, defined by a discrepancy between textual content and acoustic signal — is made possible through the combination of the discrete classifier and the PAD vector: a linguistically neutral utterance accompanied by low Valence, moderate Arousal and high Dominance produces a PAD signature incompatible with the "neutral" category, signalling to the LLM the need to adopt the *Assertive Neutrality* strategy. Top-tier models correctly handle this discrepancy in the vast majority of tested scenarios, confirming the feasibility of the approach.

7.2 Interpretation of Observed Patterns

The GPT-5 models paradox. The most counterintuitive result of the benchmark is the mid-to-lower ranking of OpenAI's GPT-5 models (gpt-5.2: 78.62; gpt-5-mini: 80.30), well below GPT-4.1 (90.3) despite being the newer generation. Metric analysis reveals that the bottleneck is *Voice Suitability* (1.87 and 2.12 respectively), which measures the suitability of generated text for speech synthesis. GPT-5 models tend to produce elaborate responses, with complex subordinate structures and argumentative transitions that read naturally but are unsuitable for synthesised speech, where brevity

and conversational rhythm are preferable. This result suggests that optimisation toward extended reasoning capability — characteristic of newer-generation models — introduces a verbosity bias that penalises voice applications. For the vocal customer service domain, GPT-4.1 therefore remains the superior choice compared to its more recent successors.

The anomalous profile of Claude Haiku 4.5. Claude Haiku 4.5 presents the most peculiar profile in the benchmark: the highest Empathy score (4.50), combined with the lowest emotion and strategy identification rates among Anthropic models (0.71/0.71). This pattern suggests a model trained to produce empathically well-calibrated responses in a manner relatively independent of the emotional diagnosis — a characteristic that may be advantageous in low-stakes scenarios (e.g. informational requests in a neutral tone), but potentially harmful in high-urgency or anger scenarios, where an empathic but strategically incorrect response risks being perceived as condescending. Deployment of Haiku 4.5 in customer service contexts should therefore be limited to domains with low emotional variance.

Size vs. architecture in open-source models. Comparison of open-source models hosted on Groq highlights that model size is not the main predictor of emotional performance. Qwen3-32B (86.2 points) outperforms Llama 3.3-70B (82.6) with nearly half the parameters, and gpt-oss-20b (84.0) outperforms Llama 3.3-70B with nearly halved latency (1.041 s vs 0.758 s). This result indicates that architectural choices and training data specific to emotional reasoning and dialogue have a greater impact than parameter scale alone.

7.3 Limitations

The present work presents several limitations that must be considered when interpreting the results.

First, the 40 gold-standard scenarios were manually constructed by a single annotator, without an inter-rater validation procedure. While manual construction ensures coverage of complex emotions, it introduces a possible subjective bias in defining the correct strategies, particularly for ambiguous emotional states such as sarcasm or latent frustration.

Second, the trained SER classifier achieves an accuracy of 66.81% on test data, a value that reflects the intrinsic difficulty of the task on heterogeneous multi-corpus data but leaves significant room for improvement. Classification errors at the SER level propagate into the JSON context provided to the LLM, potentially degrading response quality in real interactions compared to what was observed in gold-standard scenarios, where PAD values were manually set.

Third, the benchmark was conducted exclusively in Italian on simulated scenarios, not on real user interactions. Generalisability of results to other languages and to real acoustic conditions (ambient noise, regional accents, spontaneous speech) remains to be verified.

Finally, the use of Claude Opus 4.6 as the LLM-as-a-Judge introduces potential bias in the evaluation of Anthropic models, which are included in the same evaluation. While the scores obtained by Anthropic models are consistent with qualitative

expectations, validation with multiple judges or human evaluators would be desirable for future studies.

7.4 Implications for Industrial Deployment

Benchmark results provide operational guidelines for LLM model selection in a vocal customer service system based on EmotionLayer. The optimal choice depends on the specific requirements profile of the application context and can be reduced to three main scenarios.

For **high-quality** environments where latency is tolerable (e.g. specialised assistance, banking, healthcare), GPT-4.1 represents the reference choice, with the best TTS Alignment in the benchmark (4.85) and top-tier quality at contained cost.

For **high-volume** environments with tight latency constraints (e.g. e-commerce, telco), gpt-oss-120b on Groq offers the same emotional accuracy as leading models (0.90/0.90) with latency below 2 s and a cost an order of magnitude lower than equivalent proprietary models.

For environments with **limited budget** or privacy requirements excluding external APIs, gpt-oss-20b on Groq (84.0 points, 1.041 s, $0.000364) represents the best self-hosted trade-off, with performance sufficient for the majority of standard customer service scenarios.

8 Conclusions

In this work we presented **EmotionLayer**, a multimodal architecture for speech emotion recognition designed for Italian-language vocal customer service. The system addresses the empathic gap that characterises current conversational systems, which process exclusively the semantic content of speech, discarding the rich paralinguistic information that conveys the user's emotional state.

EmotionLayer integrates two neural models derived from `facebook/wav2vec2-large-xlsr-53`: a discrete Ekman emotion classifier, trained on the union of Emozionalmente, EMOVO and AI4SER (66.81% accuracy on 3,860 test samples), and a continuous PAD vector regressor, trained on AI4SER (mean CCC 0.599). The two models operate in parallel, producing a structured JSON emotional context that is provided to the downstream LLM to guide response strategy selection and speech synthesis parameters.

The systematic evaluation of 18 LLM models on 40 gold-standard scenarios via the *LLM-as-a-Judge* methodology produced three main results. First, top-tier proprietary models — Gemini 2.5 Flash Preview (91.7/100), Gemini 2.5 Pro (90.5/100) and GPT-4.1 (90.3/100) — demonstrate robust understanding of the multimodal emotional context, with emotion identification and strategy selection rates of 90–92%. Second, the open-source model **openai/gpt-oss-120b** hosted on Groq (87.1/100, 1.594 s, $0.000633) achieves the same emotional accuracy as leading models at a fraction of the cost, emerging as the optimal choice for high-volume production environments. Third, OpenAI's GPT-5 models — despite being the more recent generation — show performance below expectations in the vocal context due to anomalously low *Voice*

Suitability, signalling a potential misalignment between optimisation toward extended reasoning and the requirements of speech synthesis.

Overall, this work demonstrates that empathic voice assistants are feasible for industrial deployments in Italian, provided an architecture is adopted that explicitly integrates paralinguistic analysis into the conversational pipeline.

Future Research Directions

The results open several development directions. On the emotional recognition side, the SER classifier accuracy (66.81%) leaves ample room for improvement: the adoption of data augmentation strategies, fine-tuning on spontaneous rather than acted speech corpora, and the integration of multi-task learning techniques that simultaneously train the discrete classifier and the PAD regressor could produce significant improvements.

On the evaluation side, constructing a larger benchmark with multi-rater annotation and validation on real user interactions represent necessary steps to consolidate the ecological validity of the results. Extending the system to other languages — particularly low-resource languages for which emotional datasets are even scarcer — constitutes a further high-impact direction.

Finally, integrating EmotionLayer into end-to-end Speech-to-Speech architectures represents the most promising frontier: the ability to modulate not only the response text but directly the prosodic parameters of the synthesised voice based on the detected emotional state would open the path to genuinely empathic voice interactions, capable of adapting *how* one responds as well as *what*.

Declarations

- **Competing interests:** The author declares no financial or personal conflicts of interest that may have influenced the results or interpretation of the present work.
- **Data availability:** The datasets generated and analysed during the study, as well as the reported experimental results, are available in the public repository indicated below.
- **Code availability:** The source code required to run the *EmotionLayer* system, reproduce the experiments and verify the results is available at the following GitHub repository: EmotionLayer GitHub repository. The *Wav2Vec 2.0* models trained for emotion classification and PAD parameter regression have been published on Hugging Face and are accessible at the following links: Emotion Recognition, PAD Prediction.
- **Author contribution:** Alessio Bernardini conceived the project, designed the system architecture, implemented the models, conducted the experiments, analysed the results and wrote the manuscript.

References

[1] Ekman, P.: An argument for basic emotions. Cognition & Emotion **6**(3–4), 169–200 (1992) https://doi.org/10.1080/02699939208411068

[2] Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality **11**(3), 273–294 (1977) https://doi.org/10.1016/0092-6566(77)90037-X

[3] Decock, S.: What verbal de-escalation techniques are used in complaint handling? Journal of Pragmatics **218**, 1–15 (2023) https://doi.org/10.1016/j.pragma.2023.11.001

[4] Miller, W.R., Rollnick, S.: Motivational Interviewing: Helping People Change, 3rd edn. Guilford Press, New York (2013)

[5] Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology **76**(6), 893–910 (1999) https://doi.org/10.1037/0022-3514.76.6.893

[6] Brown, P., Levinson, S.C.: Politeness: Some Universals in Language Usage. Cambridge University Press, Cambridge (1987)

[7] Coombs, W.T.: Crisis communication: A developing field. In: Coombs, W.T., Holladay, S.J. (eds.) The Handbook of Crisis Communication, pp. 19–53. Wiley-Blackwell, Oxford (2010)

[8] Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and Semantics, Vol. 3: Speech Acts, pp. 41–58. Academic Press, New York (1975)

[9] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12449–12460 (2020)

[10] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation **42**(4), 335–359 (2008) https://doi.org/10.1007/s10579-008-9076-6

[11] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, M., Zeng, M., Yu, X., Wei, F.: WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022) https://doi.org/10.1109/JSTSP.2022.3188113 arXiv:2110.13900

[12] Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021) https://doi.org/10.1109/TASLP.2021.3122291 arXiv:2106.07447

[13] Livingstone, S.R., Russo, F.A.: 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 **13**(5), 0196391 (2018) https://doi.org/10.1371/journal.pone.0196391

[14] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing **5**(4), 377–390 (2014) https://doi.org/10.1109/TAFFC.2014.2336244

[15] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Ringeval, F., Schuller, B.W.: Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10745–10759 (2023) https://doi.org/10.1109/TPAMI.2023.3263585

[16] Costantini, G., Iaderola, I., Paoloni, A., Todisco, M.: EMOVO corpus: An Italian emotional speech database. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, pp. 3501–3504 (2014)

[17] Parada-Cabaleiro, E., Batliner, A., Schedl, M.: Emozionalmente: A Crowdsourced Italian Emotional Speech Corpus. Zenodo. Available at https://zenodo.org/record/6569930 (2021)

[18] Hugging Face Community: AI4SER: Italian Speech Emotion Dataset with PAD Annotations. Hugging Face Datasets. DOI: `10.57967/hf/6703`, Licenza CC-BY 4.0 (2024)

[19] OpenAI: GPT-4 technical report. Technical report, OpenAI (2023)

[20] Paech, S.J.: EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models (2023)

[21] Paech, S.J.: EQ-Bench 3: Emotional Intelligence Benchmark. https://github.com/EQ-bench/eqbench3 (2025)

[22] Chen, Y., et al.: EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models (2024)

[23] Arjmand, E., Heyrani Nobari, A., Rafatirad, S., Homayoun, H., Rawassizadeh, R.: Empathic grounding: Exploiting emotional dispositions for medical dialogues. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL): Industry Track, Toronto, Canada, pp. 533–545 (2023). https://doi.org/10.18653/v1/2023.acl-industry.52

[24] Lei, Y., Wen, Z., Shi, C., Zhao, J.: Empathic prompting: Improving chatbot empathy through emotional context and valence-arousal integration. Applied Sciences **14**(14), 6120 (2024) https://doi.org/10.3390/app14146120

[25] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36 (2023)

[26] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98–125 (2017) https://doi.org/10.1016/j.inffus.2017.02.003

[27] 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 (ACL), Florence, Italy, pp. 527–536 (2019). https://doi.org/10.18653/v1/P19-1050

[28] Wang, Y., Boumadane, A., Heba, A.: A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. In: Proc. Interspeech, pp. 1580–1584 (2021). https://doi.org/10.21437/Interspeech.2021-1956

[29] Siriwardhana, S., Reis, A., Weerasekera, R., Nanayakkara, S.: Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition. In: Proc. Interspeech, pp. 3755–3759 (2020). https://doi.org/10.21437/Interspeech.2020-1212

Appendix A Prompts Used for LLM Model Evaluation

This appendix reports in full the prompts used for the experimental evaluation of Large Language Models (LLMs) in the context of a multimodal Customer Care system.

The experimental architecture comprises:

- A **Generative Model (Agent Model)** tasked with analysing multimodal inputs (transcription + prosodic parameters) and producing a structured response.
- An **Evaluator Model (LLM-as-a-Judge)** tasked with verifying the adherence of the generated output to the instructions of the original System Prompt.

The prompts were kept identical for all tested models to ensure comparability, control of experimental variables, and reproducibility. Note: the prompts were intentionally written in Italian, as the system is designed for Italian-language interactions.

A.1 Prompt 1 – Multimodal Orchestrator for Empathic Customer Care

```
# ROLE
Sei un'IA avanzata specializzata in Analisi del Sentiment Multimodale e
    Customer Care Empatico.
Il tuo compito  orchestrare la risposta perfetta analizzando non solo COSA
    dice l'utente (testo), ma COME lo dice (audio/prosodia).

# INPUT DATA EXPLANATION
Riceverai un JSON contenente:
1. **vocal_sentiment**: Classificazione emotiva dall'audio.
2. **prosody (PAD Model)**:
   - **Arousal (0.0-1.0)**: Livello di energia/attivazione.
    (Basso=Calmo/Depresso, Alto=Eccitato/Arrabbiato).
   - **Valence (0.0-1.0)**: Positivit. (Bassa=Negativo, Alta=Positivo).
   - **Dominance (0.0-1.0)**: Livello di controllo percepito.
    (Basso=Sottomesso/Insicuro, Alto=Dominante/Aggressivo).
3. **transcription**: Il testo detto dall'utente.

# LOGIC GUIDELINES (Priority Rules)
Devi determinare l'emozione finale seguendo queste regole di priorit:
1. **Rilevamento Incongruenze (Sarcasmo/Ironia):** Se il 'text_sentiment'
    Positivo ma la 'prosody.valence'  Bassa e la 'prosody.dominance'  Alta ->
    L'emozione  **SARCASMO**. L'audio vince sempre sul testo.
2. **Rilevamento Urgenza:** Se 'prosody.arousal' > 0.8 -> L'emozione tende
    verso **RABBIA** (se valence bassa) o **URGENZA** (se dominance
    bassa/media).
3. **Rilevamento Passivit:** Se 'prosody.arousal' < 0.4 e 'prosody.valence' <
    0.4 -> L'emozione tende verso **INSICUREZZA** o **FRUSTRAZIONE** passiva.
```

```
# EMOTION REFERENCE TABLE (PAD Estimates)
Usa questi riferimenti approssimativi per mappare i valori di input
    all'emozione pi vicina:
- **NEUTRA**: Valence ~0.5, Arousal ~0.5
- **ENTUSIASMO**: Valence >0.7, Arousal >0.7
- **SODDISFAZIONE**: Valence >0.6, Arousal <0.5
- **INCERTEZZA**: Valence <0.5, Arousal ~0.5, Dominance <0.4
- **RABBIA**: Valence <0.3, Arousal >0.8, Dominance >0.7
- **FRUSTRAZIONE**: Valence <0.4, Arousal >0.6, Dominance <0.5
- **SARCASMO**: Discrepanza Testo/Audio (vedi Logic Guidelines)
- **URGENZA**: Valence <0.4, Arousal >0.8, Dominance <0.5
- **INSICUREZZA**: Valence <0.4, Arousal <0.4, Dominance <0.3

# STRATEGIES
Scegli la strategia basandoti sull'emozione rilevata:
1. DIRECT EFFICIENCY (Neutra): Risoluzione rapida. Fornisci informazioni
    precise, strutturate e proattive. Velocit standard (1.0), tono
    informativo, pitch bilanciato.
2. DE-ESCALATION ATTIVA (Rabbia/Frustrazione): Priorit alla calma. Ridurre la
    velocit (0.8) e il pitch (low). Evitare giustificazioni; usare un tono
    'apologetic' e offrire una soluzione immediata o un escalation path.
3. EMPATHIC VALIDATION (Insicurezza/Incertezza): Rallentare il ritmo (0.85),
    tono 'empathetic'. Validare verbalmente l'ostacolo prima di fornire
    istruzioni passo-passo.
4. POSITIVE MIRRORING (Entusiasmo/Soddisfazione): Aumentare energia (speed
    1.1, pitch high), tono 'cheerful'. Ringraziare per il feedback e
    consolidare la loyalty.
5. ASSERTIVE NEUTRALITY (Sarcasmo): Tono 'firm' o 'professional', pitch low,
    velocit standard. Rispondere esclusivamente ai fatti tecnici.
6. RAPID RESOLUTION (Urgenza): Massima efficienza (speed 1.25), tono
    'professional', pitch default. Messaggi brevi, orientati all'azione
    immediata.

# OUTPUT FORMAT
Restituisci ESCLUSIVAMENTE un oggetto JSON valido. Non aggiungere markdown
    (```json), non aggiungere testo introduttivo.

{
    "reasoning": "...",
    "selected_strategy": "STRATEGIA_SCELTA",
    "emotion_detected": "EMOZIONE_RILEVATA",
    "tts_config": {
        "speed": [0.8-1.2],
        "tone": "string",
        "pitch": "string"
    },
    "response_text": "Il testo della risposta."
}
```

A.2 Prompt 2 – Evaluation Judge for Voice UX and AI Evaluation

```
Sei un Giudice esperto in Voice UX e AI Evaluation.
Il tuo compito  valutare quanto bene un modello LLM ha seguito le istruzioni
    specifiche del suo System Prompt originale.

Analizza l'input dell'utente (trascrizione + emozione) e l'output del modello
    per assegnare i voti.

METRICHE DI VALUTAZIONE:

1. EMOTION_MATCH (0 o 1):
   - 1: Il modello ha predetto ESATTAMENTE l'emozione che ci aspettavamo.
   - 0: Il modello ha predetto un emozione totalmente sbagliata da quella che
    ci aspettavamo.

2. STRATEGY_MATCH (0 o 1):
   - 1: Il modello ha scelto ESATTAMENTE la strategia richiesta dal prompt
    originale per quella emozione.
   - 0: Ha scelto una strategia sbagliata o non in lista.

3. RELEVANCE (1-5):
   - 1: Allucinazione totale o risposta fuori tema.
   - 5: Risposta perfettamente pertinente alla richiesta dell'utente.

4. TTS_ALIGNMENT (1-5):
   - Valuta se i parametri 'tts_config' rispettano le regole del prompt
    originale.
   - 1: Parametri opposti (es. speed 1.2 per rabbia).
   - 5: Parametri perfettamente allineati alle istruzioni.

5. VOICE_SUITABILITY (1-5):
   - Il testo 'response_text'  adatto per essere letto ad alta voce?
   - 1: Troppo lungo, linguaggio burocratico, formattazione non leggibile
    (markdown, elenchi).
   - 5: Breve, frasi semplici, naturale, niente elenchi puntati o caratteri
    speciali.

6. EMPATIC_RESPONSE (1-5):
   Dai una valutazione da 1 a 5 in base a quanto la risposta 'response_text'
    sia empatica.
   - 1: Risposta per nulla empatica.
   - 5: Risposta Empatica.

Rispondi SOLO con questo JSON:
{{
   "emotion_match": 0/1
   "strategy_match": 0/1,
```

```
   "relevance_score": 1-5,
   "tts_score": 1-5,
   "voice_suitability_score": 1-5,
   "empatic_response": 1-5
   "explanation": "..."
}}
```

Appendix B Tested Model Pricing

Table B1 reports the input and output token costs (as declared at the time of experimentation) for each model used in the tests. Prices are expressed in USD per million tokens.

Model Name	Provider	Input ($/M)	Output ($/M)
gpt-5-mini-2025-08-07	OpenAI	0.25	2.00
gpt-5.2-2025-12-11	OpenAI	1.75	14.00
gpt-4.1-2025-04-14	OpenAI	2.00	8.00
gpt-4.1-mini	OpenAI	0.40	1.60
Qwen/Qwen2.5-7B-Instruct:together	HuggingFace	0.30	0.30
Qwen/Qwen2.5-72B-Instruct:novita	HuggingFace	0.38	0.40
mistralai/Mistral-7B-Instruct-v0.2:featherless-ai	HuggingFace	0.25	0.25
openai/gpt-oss-120b	Groq	0.15	0.60
openai/gpt-oss-20b	Groq	0.075	0.30
llama-3.1-8b-instant	Groq	0.05	0.80
llama-3.3-70b-versatile	Groq	0.59	0.79
qwen/qwen3-32b	Groq	0.29	0.59
claude-sonnet-4-5	Anthropic	3.00	15.00
claude-haiku-4-5	Anthropic	1.00	5.00
claude-opus-4-5	Anthropic	5.00	25.00
gemini-2.5-flash-preview-09-2025	Gemini	0.30	2.50
gemini-2.5-pro	Gemini	1.25	10.00
gemini-2.5-flash-lite	Gemini	0.10	0.40

Table B1: Pricing of tested LLM models (USD per million tokens).