

©Copyright 2025

Raviteja Tanikella

Emotionally Intelligent Voice Language Models For Mental Health Therapy

Raviteja Tanikella

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Computer Science and Software Engineering

University of Washington

2025

Committee:

Dong Si

Annuska Zolyomi

Afra Mashhadi

Program Authorized to Offer Degree:
Computer Science and Software Engineering

University of Washington

Abstract

Emotionally Intelligent Voice Language Models
For Mental Health Therapy

Raviteja Tanikella

Chair of the Supervisory Committee:
Dong Si
Computing and Software Systems

Conversational assistants for mental health therapy primarily rely on text-based models or cascaded architectures that first transcribe speech to text, a process that discards crucial paralinguistic information. This information bottleneck limits the AI’s ability to perceive critical emotional cues and broader psychological states, hindering its capacity to reason over human voice and provide effective therapy. This thesis details the development of an emotionally intelligent voice language model designed to overcome these limitations. The process began with a systematic evaluation of heuristic-based approaches and end-to-end model architectures, where automated benchmarking and human study confirmed that large audio language models provided the most effective foundation for voice understanding and reasoning. Building on these findings, I propose and implement policy optimization methods to fine-tune Qwen2.5-Omni on therapy data. The resulting aligned model demonstrated improved performance in benchmark evaluations, exhibiting emotional intelligence and generating therapeutically relevant responses. By presenting a complete development framework, from architectural validation to targeted alignment, this research establishes a clear and proven roadmap for creating the next generation of efficient and adaptive voice language models for mental health therapy and multimodal conversational AI.

TABLE OF CONTENTS

	Page
List of Figures	v
List of Tables	vii
Glossary	viii
Chapter 1: Introduction	1
1.1 Motivation	1
1.2 Importance of Voice Modality	3
1.3 Problem Statement	4
1.4 The iCare Project	5
Chapter 2: Related Works	7
2.1 Emotional Intelligence in Conversational AI	7
2.2 Personalized and Adaptive Architectures	8
2.3 Expressive Text-to-Voice Synthesis and Voice Presence	8
2.4 From Speech Emotion to Voice Emotion Recognition	9
2.5 Speech Language Models	10
2.6 Joint Audio and Speech Understanding	10
Chapter 3: Methodology	13
3.1 Overview of Research Framework	13
3.2 Analysis of End-to-End LALM Training	13
3.2.1 Architectural Design	13
3.2.2 fine-tuning the Reasoning Module on Mental Health Data	15
3.2.3 Automated Metric-Based Evaluation	18
3.2.4 Audio Instruction Tuning Challenges and Methodological Pivot	22

3.3	Heuristic-Based LALM with Discretized Audio Features	23
3.3.1	Initial Techniques and Their Limitations	24
3.3.2	Heuristic Discretization via Semantic Vector Quantization	25
3.3.3	Heuristic Feature Definition	27
3.3.4	Does These Features Contribute to Emotional Intelligence?	30
3.3.5	Empirical Validation of Heuristic Thresholds	31
3.3.6	Pros & Cons of Heuristic-Based Approach	33
3.4	From Heuristics to Native Voice Understanding	34
3.5	Comparative Benchmarking and Model Selection	36
3.5.1	Benchmark Dataset Preparation	36
3.5.2	Models Under Evaluation	37
3.5.3	Evaluation Framework: LLM-as-a-Judge	38
3.5.4	Performance Verification on VoiLa-Bench	39
3.5.5	Validation of Voice Language Model Based on VocalBench	40
3.6	User-Based Study for Model Selection	41
3.6.1	Ethical Considerations and IRB Approval	41
3.6.2	Study Design and Procedure	41
3.6.3	Evaluation Method	42
3.7	Alignment of Voice Language Model for Therapy	43
3.7.1	Limitations For Post-training	43
3.7.2	Dataset Preparation for Alignment	44
3.7.3	Alignment via Direct Preference Optimization	45
3.7.4	Implementation and Evaluation of DPO Alignment	46
3.7.5	Human Expert Evaluation of DPO Model	47
Chapter 4:	Experiments & Results	50
4.1	Supervised Fine-tuning of Reasoning Module in LALM	50
4.1.1	Analysis of Semantic Equivalence (BERTScore)	50
4.1.2	Analysis of Lexical Precision (BLEU Score)	51
4.1.3	Analysis of Lexical Recall (ROUGE Score)	53
4.1.4	Analysis of Linguistic Quality Metrics (METEOR & Distinct)	54
4.1.5	Conclusion and Final Model Selection	55
4.2	Validation of Heuristic Thresholds	57

4.3	Comparative Benchmarking and Model Selection	60
4.3.1	Benchmarking on Prepared DAIC-WOZ Dataset	60
4.3.2	Benchmarking on Voila-Bench	66
4.3.3	Analysis of VocalBench Results	70
4.3.4	User Survey Analysis	72
4.4	Evaluation of DPO Model	75
4.4.1	Metric-Based Evaluation	75
4.4.2	Human Evaluation	80
Chapter 5:	Conclusion	84
Chapter 6:	Future Work	87
6.1	Scarcity of High-Quality Mental Health Data	87
6.2	Rigorous Domain Expert Evaluations	88
6.3	Comparing Qwen2.5-Omni Model on Diverse Audio Lengths	88
6.4	Supervised Fine-Tuning vs. Direct Preference Optimization	89
6.5	State-Space Models (SSMs) as a Promising Future	89
Bibliography		91
Appendix A:	Analysis of State Space Models for Voice Language Modelling	105
A.1	Training Speed, Time, and Memory	105
A.2	Model Weights: Performance–Parameter Ratio	107
A.3	Long-Duration Audio Processing	109
A.4	Capturing Long and Short Range Dependencies	110
A.5	Context-Length Scaling	111
A.6	On-Device Implementations	113
Appendix B:	Supplementary Materials	115
B.1	Fine-Tuning Hyper-parameters & Scripts	115
B.1.1	SFT Configuration for Stage-1 Fine-Tuning	115
B.1.2	DPO Training Configuration for Qwen2.5-Omni	116
B.1.3	Finetuning and Voice Generation Scripts	116
B.2	Benchmarking and Evaluation Prompts	125

B.2.1	Model Prompt for DAIC-WOZ Benchmark	125
B.2.2	Gemini 2.5 Pro Judge Prompt for DAIC-WOZ	127
B.2.3	Gemini 2.5 Pro Judge Prompt for Voila Bench	128
B.3	Additional Code and Datasets	129

LIST OF FIGURES

Figure Number	Page
1.1 Cascaded architecture of ASR + LLM models.	2
1.2 Cascaded vs Voice-Language Model Architecture.	4
1.3 A sample conversation with Carebot (iCare).	5
3.1 LALM Architecture consisting of perception and reasoning modules.	14
3.2 Conversion of audio into semantic descriptors using Heuristic Translator . . .	29
4.1 Mean Average scores across all the three Llama 3.1 models	56
4.2 Mean Average scores comparing MentalChat-SFT and Gemini-2.5-Pro	56
4.3 Correlation of scores for the vocal energy heuristic indicating positive trend.	58
4.4 Correlation of scores for the pitch variability heuristic indicating positive trend.	59
4.5 Comparision of all the three runs on DAIC-WOZ dataset	60
4.6 ESHRO Comparision of all the three runs on DAIC-WOZ dataset	64
4.7 Overall model performance comparison on the VoilaBench benchmark	67
4.8 Score distributions for Voila, Qwen2.5-Omni, and Qwen2-Audio on the VoilaBench benchmark	69
4.9 Average quality ratings on a 1-4 scale for Model A (Qwen2.5-Omni) and Model B (Voila-Chat) across four key criteria.	72
4.10 Average user ratings (1-10) per audio input for Model A (Qwen2.5-Omni) and Model B (Voila-Chat)	74
4.11 Mean BERT Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.	76
4.12 Mean BLEU Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.	77
4.13 Mean ROUGE Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.	78
4.14 Mean METEOR Scores Comparison. The DPO-aligned model shows a significant improvement in semantic alignment.	79

4.15	Mean Perplexity Scores Comparison. Lower perplexity for the DPO model indicates improved fluency and model confidence.	80
4.16	Mean Distinct Scores Comparison. The DPO model shows a notable increase in bigram diversity (Distinct-2), indicating more varied phrasing.	81

LIST OF TABLES

Table Number	Page
3.1 A summary of the pros and cons of using a heuristic-based approach for par- alinguistic analysis.	33
4.1 Mean BERT scores for each model. Best scores are in bold.	50
4.2 Mean BLEU scores (n=1-4). Best scores are in bold.	52
4.3 Mean ROUGE scores for each model. Best scores are in bold.	53
4.4 Mean METEOR and Distinct scores for each model.	54
4.5 Frequency of Ratings (1-10) for Each Response Type Across All Runs	62
4.6 Evaluation performance of 12 models on VocalBench. Bold indicates the op- timal result in each subgroup. Source: (Liu et al., 2025).	71
4.7 Human Expert Model Preferences across Evaluation Criteria showing check marks indicating better performance for each respective model.	82
B.1 Complete fine-tuning hyper-parameters for SFT LoRA on Llama-3.1-8B-Instruct	115
B.2 Complete DPO Alignment Hyperparameters for Qwen2.5-Omni-7B	116

GLOSSARY

BEHAVIORAL HEALTH SERVICES: Healthcare services focused on the diagnosis and treatment of mental health and substance use disorders.

CLINICAL INTERVIEWS: Structured or semi-structured conversations conducted by healthcare professionals to gather comprehensive information for diagnosis or treatment.

CONVERSATIONAL AI ASSISTANTS: Artificial Intelligence (AI) systems designed to interact with users through natural language conversations.

DIRECT PREFERENCE OPTIMIZATION (DPO): An alignment method for LLMs that uses the model’s own probabilities to optimize responses without a separate reward model.

DOMAIN ADAPTATION: The process of fine-tuning a machine learning model to perform well on data from a specific domain, different from its original training data.

EMOTIONAL INTELLIGENCE (EI): The ability to perceive, understand, reason about, and express emotions effectively.

FEATURE DISCRETIZATION: The process of converting continuous acoustic feature values into discrete semantic tiers (e.g., very low to very high).

HEURISTIC-BASED APPROACH: An approach that employs predefined heuristic rules to summarize and encode vocal acoustic features into semantic descriptors for language models.

INSTRUCTION TUNING (IT): A fine-tuning method where models are trained to follow explicit instructions, improving task-specific responses.

LARGE LANGUAGE MODEL (LLM): A neural network model with a large number of parameters, trained on extensive text data to understand, generate, and reason about human language in a coherent and context-aware manner.

LARGE AUDIO LANGUAGE MODEL (LALM): A model designed to jointly understand audio & speech inputs, and generate language (text).

LOW-RANK ADAPTATION (LORA): A technique for computationally efficient fine-tuning of large language models by adding low-rank matrices to pre-trained weights.

LTU-AS: Architectural design inspired by a model combining acoustic perception with language reasoning capabilities.

MULTIMODAL LARGE LANGUAGE MODELS (MLLMs): Models that can process, understand, and respond to information from multiple data types, such as text, images, and audio.

PARALINGUISTIC CUES: Non-verbal elements of speech such as tone, pitch, loudness, and rhythm that conveys emotional or contextual meaning.

PITCH VARIABILITY: The degree of fluctuation in the pitch of speech, serving as an indicator of prosody and emotional state.

SEMANTIC VECTOR QUANTIZATION: A technique that assigns continuous acoustic features to discrete codebook vectors representing semantic tiers for easier integration with language models.

SPEECH RHYTHM: The perceived pace or cadence of speech, reflecting the timing and flow of spoken language.

STATE-SPACE MODELS (SSMs): Models that describe a system using a set of input, output, and state variables, often used for modeling dynamic systems and enabling efficient processing of long sequences.

SUPERVISED FINE-TUNING (SFT): The process of refining a pre-trained model on labeled data to enhance performance on a specific task.

TEXT-TO-SPEECH (TTS) SYSTEMS: Models that convert written text into speech.

VOCAL BIOMARKERS: Specific vocal features that can indicate physical or mental health conditions.

VOCAL ENERGY: The intensity or loudness of voice signals, commonly measured by the root mean square (RMS) of the audio waveform.

VOICE LANGUAGE MODELS (VOLMs): Models designed to natively understand and reason over speech and the emotional cues in human voices, and to generate responses in a natural, human-like voice.

ACKNOWLEDGMENTS

I want to express my deepest gratitude to Dr. Dong Si, my committee chair, whose guidance and encouragement shaped this project from inception to completion. Dr. Si's support was crucial in giving me the freedom to follow my interests and conduct impactful, applicable research. I am equally thankful to my committee members, Dr. Annuska Zolyomi and Dr. Afra Mashhadi, for their valuable insights and support throughout the research process. Their feedback was particularly crucial in transforming a strong technical model into a therapeutically relevant one.

My heartfelt thanks go to individuals whose direct support was invaluable. Specifically, I am profoundly grateful to Anvitha Lolla and Srikanth Kotikalapudi. Their constant support and dedicated assistance throughout my thesis journey, particularly during the crucial human evaluation phases, played a pivotal role in the real-world validation of my models and methods. Their contributions made a tangible difference to this research.

I also want to acknowledge Surbhi Gupta and Sugam Jaiswal. It was great working alongside them, and their collaborative spirit was instrumental in conducting publishable research that significantly contributed to this thesis. Special thanks to Neeraj Baipureddy for crucial support throughout my thesis journey.

I am especially grateful to Dr. Saurabhchand Bhati, Post-Doctoral Fellow at MIT, for his thoughtful advice and mentorship. Our talks and discussions provided me with new perspectives that proved to be crucial contributions to this thesis.

My deepest gratitude extends to my family and friends. Your constant support, endless encouragement, and understanding have been a continuous source of strength

throughout this sometimes challenging, yet immensely rewarding, academic journey.

Beyond these direct collaborations, I extend my appreciation to all members of the iCare group. Their collective knowledge and support provided an enriching environment throughout my research journey.

I also wish to acknowledge the broader community of developers and maintainers of various datasets and open-source tools. Their invaluable work provided the essential foundations for the empirical evaluations and modeling efforts, truly making much of this research possible.

I owe sincere thanks to the UW IT Department, and especially to Rene Gomez, for his exceptional assistance and dedication. His efforts in maintaining and swiftly resolving resource-related issues made the entire process seamless, allowing me to focus on my research without interruption.

Finally, I am sincerely grateful to the University of Washington for providing an encouraging academic environment and access to the resources that made this research possible. I also want to express special thanks to my graduate advisor, Jerry Vasquez, for his invaluable guidance throughout my thesis journey, ensuring all academic commitments were met. His support was incredibly helpful.

To everyone who has contributed, offered encouragement, or inspired me along the way, I am profoundly grateful.

DEDICATION

To my parents, whose unwavering love, support, and sacrifices made this journey possible.

This achievement is as much theirs as it is mine.

Chapter 1

INTRODUCTION

1.1 Motivation

The surging prevalence of mental health conditions has become a growing concern worldwide, with anxiety, depression, and related disorders now recognized as major contributors to global disease burden (World Health Organization, 2023). Traditional models of mental health care are often unable to address this demand efficiently due to clinician shortages, geographical barriers, financial costs, and persistent social stigma surrounding therapy (Patel et al., 2018; Kazdin and Rabbitt, 2013). These challenges have catalyzed demand for scalable, accessible solutions that can reach populations underserved by existing systems. Artificial Intelligence (AI) has emerged as a promising avenue to address these limitations, particularly through the development of conversational assistants for mental health support (Inkster et al., 2018). Such AI-driven interventions have the potential to offer low-barrier, on-demand, and confidential guidance, education, and therapeutic support, positioning them as powerful complements to traditional care.

However, most existing conversational assistants developed for mental health support function solely through text, engaging users via chat-based interfaces or text messaging (Miner et al., 2016). Influential early systems, such as Woebot, demonstrated the potential of delivering structured Cognitive Behavioral Therapy (CBT) through automated, text-only dialogues, showing efficacy in reducing symptoms of depression and anxiety (Fitzpatrick et al., 2017). This text-first paradigm was driven by the maturity of natural language processing for text, user privacy considerations, and the desire to create a discreet and accessible entry point for mental health support (Provoost et al., 2017). Yet, the desire for more naturalistic, convenient, and human-like interaction, mirroring the widespread adoption

of voice commands in consumer technology, has motivated a necessary evolution toward speech-based interfaces (Ma et al., 2024). The most direct engineering path to enable voice interaction is a cascaded pipeline architecture, which modularly combines a high-performance Automatic Speech Recognition (ASR) model with a separate Large Language Model (LLM) (see Figure 1.1). In this configuration, user speech is first transcribed into text by the ASR module, and this transcript is then processed by the LLM to generate a response (Nguyen et al., 2024), and might use a Text-to-Speech (TTS) model to synthesize the response into audio (Aleksic and Zhou, 2025).

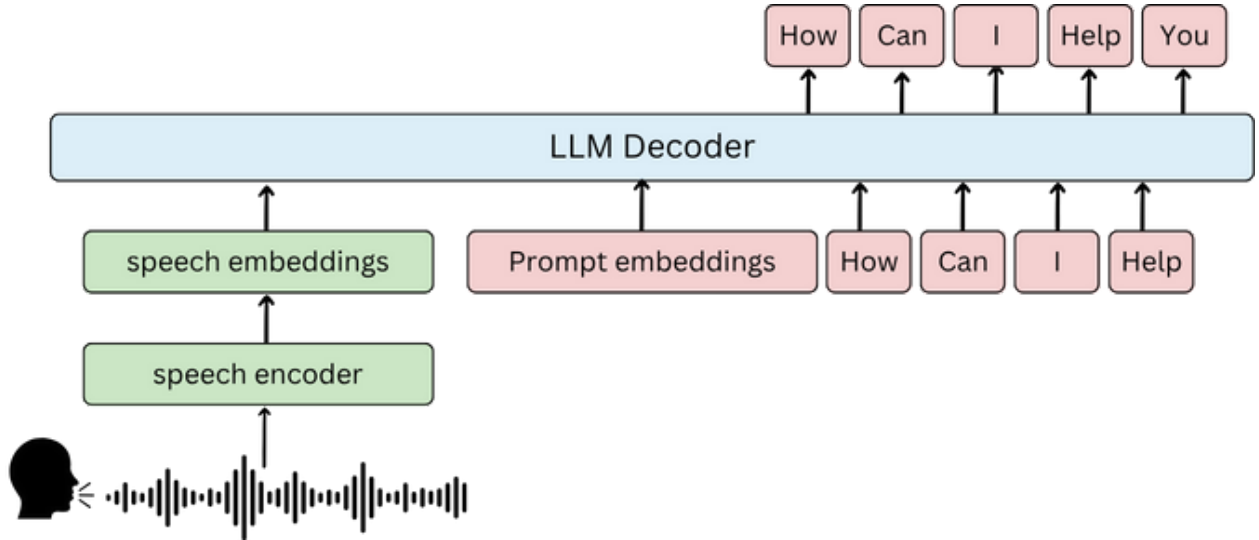


Figure 1.1: Cascaded architecture of ASR + LLM models.

While these cascaded approaches expand access to speech interaction, they rest on the flawed assumption that a textual transcript captures all relevant information present in a therapeutic conversation. Spoken communication conveys not only the explicit content of words but also rich paralinguistic cues—features such as prosody, pitch, tone, speaking rate, and pauses—that carry a substantial portion of emotional and psychological information (Cummins et al., 2015; Scherer, 2003). The practice of reducing voice to text strips away this rich channel, effectively creating an information bottleneck. As a consequence, crucial

elements indicating distress, improvement, or urgency are often missed, and model outputs become insensitive to shifts in affect or engagement levels that a human therapist would naturally perceive.

1.2 Importance of Voice Modality

The loss of paralinguistic information is particularly detrimental in mental health contexts. Spoken language conveys critical affective data through vocal cues such as intonation, pace, and timbre, which are essential for accurately interpreting a user’s psychological state but are lost in text transcripts (Low et al., 2020; Vanello et al., 2012). Neglecting these cues hinders an AI’s ability to respond with genuine empathy and limits its utility even in clinical support tools, where vocal biomarkers could otherwise offer diagnostic insights (Krishna et al., 2025). Furthermore, as the most natural mode of human interaction, voice is fundamental to the future of conversational AI. It enables immediate, hands-free interfaces, which are crucial for accessibility, particularly for users with motor or speech impairments (Chen et al., 2024a; Gnani.ai, 2025). Advanced speech-to-speech models are now capable of handling complex conversational dynamics, making interactions feel more natural and effective than text-based alternatives (Andrade, 2025). Therefore, to create AI assistants that are truly context-aware, empathetic, and intelligent, it is necessary to develop models that can process the complete vocal signal along with speech. This thesis directly addresses this need by developing and validating an emotionally intelligent voice language model designed to overcome the information bottleneck of text-centric and cascaded systems.

Voice is a rich source of emotional data and contains vocal biomarkers that are critical for therapeutic applications and diagnostics in healthcare (Nature, 2025). The analysis of the voice using AI has shown potential to significantly improve diagnostic accuracy when compared with traditional methods, such as assessments based on demographic data or clinical interviews (Canary Speech, 2024; Behavioral Health Tech, 2023). AI models can detect these vocal biomarkers with high precision, in some cases achieving accuracy rates of over 90% for specific conditions like Parkinson’s disease or dementia (EMJ, 2024; Shigeyoshi

et al., 2025). Although voice-based AI agents can offer a more natural and, in some contexts, more effective user experience, they are less common than their text-based counterparts, which are often preferred for interactions requiring detailed, reviewable information or in environments where silence is necessary (Teneo.ai, 2025; Reddit Community, 2025).

1.3 Problem Statement

As shown in Figure 1.2, the central thesis objective, or the problem statement of this thesis, is to answer the following:

“How can we develop an emotionally intelligent voice language model that moves beyond cascaded architectures by directly processing raw audio, enabling it to reason over the complete vocal signal and apply domain-specific knowledge to generate therapeutically relevant responses?”

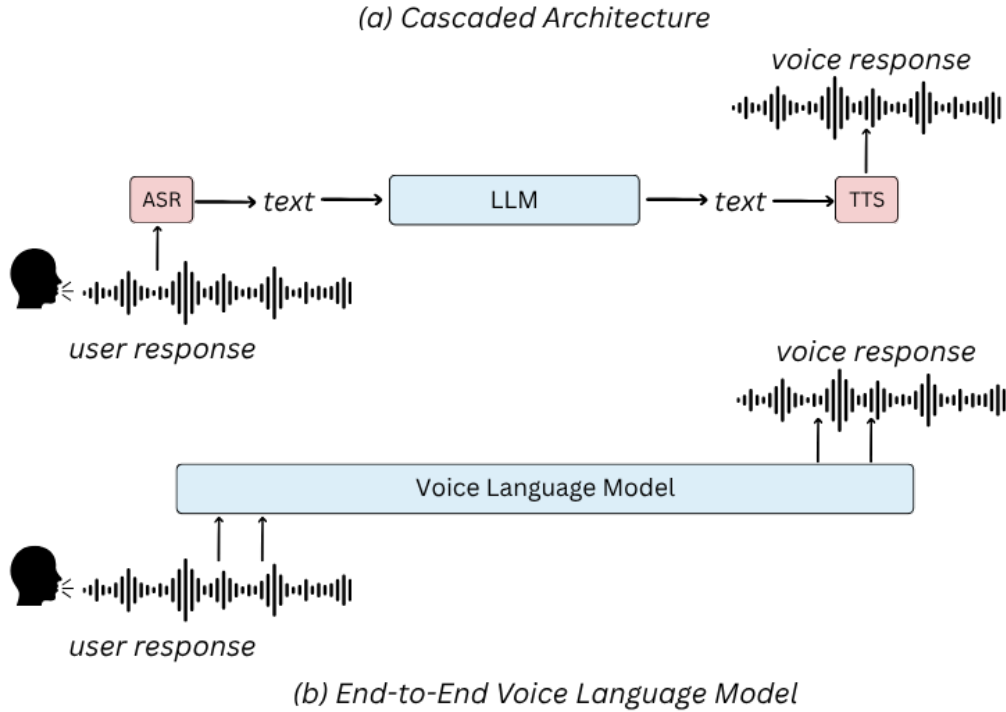


Figure 1.2: Cascaded vs Voice-Language Model Architecture.

1.4 The iCare Project

This thesis builds upon the foundational work of the iCare Project, an initiative by the Data Analysis and Intelligent Systems research group at the University of Washington aimed at addressing the global mental health crisis through AI-powered conversational agents. The project’s goal is to bridge critical gaps in mental healthcare by developing digital assistants that offer empathetic support to individuals with mental health conditions and their caregivers. The project has produced two flagship text-based systems: Carebot, which directly engages users with empathetic, story-driven interactions, and CoachBot, which serves as a virtual coach for caregivers, using the EPITOME framework to enhance their empathetic communication skills. Both systems have evolved architecturally, leveraging powerful LLMs such as GPT-3 and, more recently, Llama2 to improve the quality, adaptability, and scalability of their responses (Naswa et al., 2024; Mavila et al., 2024). A sample conversation with the Carebot is shown in Figure 1.3.

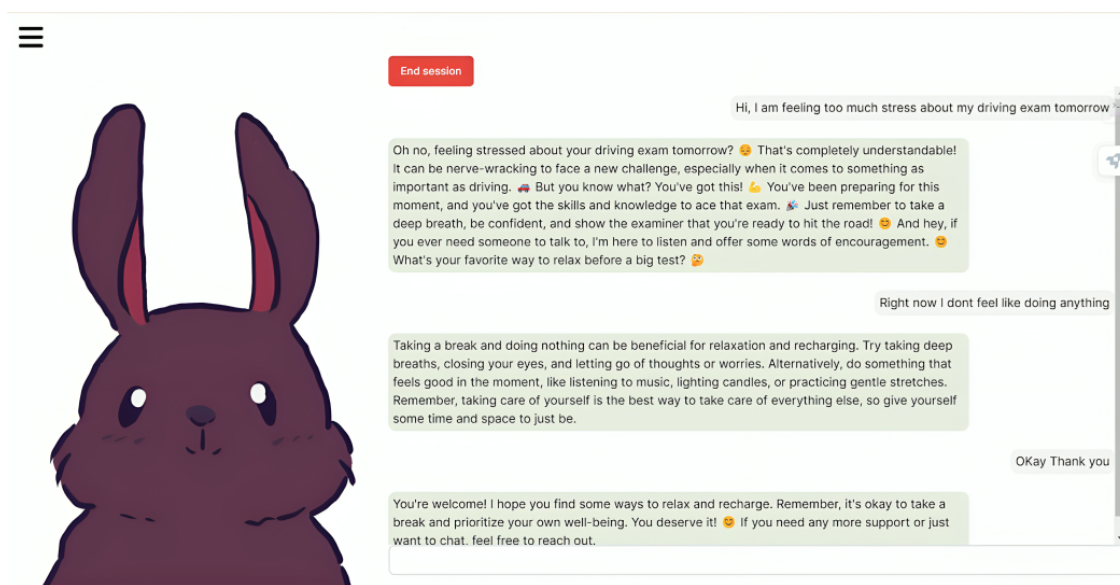


Figure 1.3: A sample conversation with Carebot (iCare).

While the iCare Project has established the efficacy of LLM-based therapeutic chatbots,

the next frontier in this research involves moving beyond one-size-fits-all text interactions toward more personalized and adaptive systems. This aligns with the emerging research direction of Personality Adaptive Conversational AI (PACA), recently presented at the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB 2024). The core principle of PACA is to create a conversational assistant that can dynamically adapt its communication style to a user's unique personality traits and emotional states, thereby enhancing the therapeutic alliance and user engagement (Jaiswal et al., 2024a).

Chapter 2

RELATED WORKS

2.1 Emotional Intelligence in Conversational AI

The capacity for Emotional Intelligence (EI), defined as the ability to perceive, understand, manage, and utilize emotions to guide thought and action, is a cornerstone of effective human interaction (Salovey and Mayer, 1990). In the context of AI, this concept is often operationalized through the field of affective computing, which aims to develop systems that can recognize, interpret, process, and simulate human emotions (Picard, 1995). For a conversational agent to be truly effective, particularly in sensitive domains like mental health, it must move beyond mere linguistic competence and embody principles of EI. This requires an architecture capable of processing the subtle emotional and psychological cues that are often conveyed non-verbally.

Early efforts to instill EI in AI focused primarily on text-based sentiment analysis. However, recent advancements have pushed beyond simple positive-negative classification. Current research now evaluates the ability of Large Language Models (LLMs) to solve complex, performance-based emotional intelligence tests, demonstrating a growing proficiency in understanding nuanced emotional scenarios from text alone (Zhang et al., 2025). Despite this progress, text as a modality remains fundamentally limited. To capture a more complete emotional picture, the field is increasingly turning to multimodal approaches. State-of-the-art models now integrate various data streams—such as text, voice, and facial expressions—to achieve a more comprehensive understanding of human emotion (Li et al., 2025b; Triantafyllopoulos et al., 2024). The development of rigorous benchmarks, such as EmoBench-M, is designed specifically to evaluate the EI capabilities of Multimodal Large Language Models (MLLMs) across complex social and conversational scenarios (Hu et al., 2025). This research

trend underscores a broad consensus: achieving robust emotional intelligence requires processing the full spectrum of human communicative signals, moving beyond what text can capture (Triantafyllopoulos et al., 2024).

2.2 Personalized and Adaptive Architectures

A key strategy for developing advanced conversational assistants is personalization. The research direction of PACA seeks to create assistants that can dynamically adapt their conversational style to a user’s unique personality traits and emotional states. As demonstrated in our prior work, this can be achieved by fine-tuning an LLM, such as Llama2, on domain-specific data and feeding it with extracted personality scores to create a system that is adaptive and therapeutically effective (Touvron et al., 2023; Jaiswal et al., 2024a). While PACA represents a significant step forward in personalizing text-based interactions, its efficacy is still constrained by the input modality; true personality and emotional adaptation require access to the richer, more immediate data present in the human voice.

2.3 Expressive Text-to-Voice Synthesis and Voice Presence

The limitations of traditional TTS models, which often produce robotic and monotonous audio, have motivated a recent shift toward expressive Text-to-Voice (TTV) models. The primary goal of this emerging research is to achieve “voice presence”—the quality that makes a synthesized voice feel natural, responsive, and emotionally aware in a live conversation (Sesame, 2025). Unlike cascaded pipelines that separate text generation from audio synthesis, these new models often use end-to-end architectures to generate not just words, but also the rich prosodic and affective cues that define human-like speech.

Several commercial and research entities are at the forefront of this effort. Sesame AI has developed a Conversational Speech Model (CSM) that processes text and audio context within a unified framework, using a dual-token strategy to capture both linguistic content (semantic tokens) and detailed voice characteristics like timbre and pitch (acoustic tokens)

(Kuka, 2025). This allows the model to implicitly learn conversational dynamics, such as natural pauses and hesitations, leading to more realistic speech (Techjays, 2025). Hume AI has introduced Octave, which it describes as a “speech-language model” rather than a traditional TTS system (Hume AI, 2025b). Octave is designed to understand the emotional intent behind text and can take natural language instructions, such as “whisper fearfully,” to guide its vocal delivery, offering creators granular control over the output’s affective quality (Hume AI, 2025a). Similarly, ElevenLabs has focused on enhancing expressiveness in its models, with recent versions incorporating features for multi-speaker dialogue and the ability to generate non-verbal vocalizations like sighs and laughter directly from textual cues (ElevenLabs, 2025).

While these state-of-the-art TTV models represent a significant leap forward in generating human-like speech, their value extends to critical applications such as synthetic data generation for training more robust conversational systems, a topic explored further in our methodology. However, their primary innovation lies in the output or synthesis stage of a conversation. They excel at creating an expressive persona for the AI but do not inherently address the core challenge of understanding and reasoning over the emotional and paralinguistic content of the user’s input. This distinction is critical; creating an assistant that sounds empathetic is a different and equally important problem than creating one that can truly perceive and react to the user’s vocal emotions.

2.4 From Speech Emotion to Voice Emotion Recognition

Within the domain of affective computing, it is important to distinguish between two related but distinct tasks. Speech Emotion Recognition (SER) traditionally focuses on classifying a speaker’s emotional state by analyzing acoustic features directly tied to the linguistic content of their speech (Schuller and Batliner, 2013). In contrast, a more holistic approach, which can be termed Voice Emotion Recognition (VER), aims to interpret emotion from the entire vocal signal. This includes not only the spoken words but also crucial non-speech vocalizations such as sighs, laughter, hesitations, and changes in breathing patterns.

2.5 Speech Language Models

A significant research effort has focused on developing Speech Language Models (SLMs) that operate in a direct speech-to-speech paradigm, bypassing the traditional ASR-LLM-TTS pipeline. This approach aims to reduce latency and generate more expressive, human-like audio. Foundational work like Google’s Translatotron demonstrated the feasibility of direct speech-to-speech translation (Jia et al., 2019). This has evolved into state-of-the-art generative models like Meta’s Voicebox and Microsoft’s VALL-E, which excel at controllable, high-fidelity audio synthesis from minimal vocal prompts (Le et al., 2023; Wang et al., 2023). More advanced hybrid models, such as Moshi and SpeechGPT, further refine this paradigm by incorporating an internal text-based reasoning step to improve the linguistic coherence and interactivity of their spoken responses (Défossez et al., 2024; Zhang et al., 2023).

While these models represent the frontier of realistic speech generation, their architectural focus is fundamentally oriented toward the output side of a conversation. Their internal mechanisms are optimized to produce a high-quality spoken response, not to perform deep analytical reasoning on the nuances of the input signal. The key distinction is this: SLMs are designed to solve the problem of how an AI assistant sounds, which is a critical challenge for creating natural dialogue. In contrast, the work in this thesis addresses the different, but equally important, problem of how an AI assistant listens and understands. We focus on the input side, developing a model architected for the explicit purpose of interpreting the full spectrum of a user’s vocal signal for applications requiring deep emotional and psychological insight.

2.6 Joint Audio and Speech Understanding

The limitations of both cascaded systems and traditional SER have motivated a recent shift towards end-to-end models capable of joint audio and speech understanding. This paradigm shift was arguably catalyzed by the introduction of models that could “Listen, Think, and Understand” both audio and speech within a unified framework. A pivotal work in this

domain is the LTU-AS model, which effectively established the blueprint for what can be termed as Large Audio Language Models (LALMs) (Gong et al., 2023b).

The architectural innovation of LTU-AS was to integrate a pre-trained, powerful audio encoder directly with a pre-trained LLM. Specifically, the authors utilized the encoder from OpenAI’s Whisper model to convert raw audio input into a sequence of discrete acoustic tokens (Radford et al., 2023). These tokens, which represent the acoustic properties of the sound, were then prepended to a textual instruction and fed directly into the LLaMA language model. This approach bypassed the need for a separate text-transcription step entirely, allowing the LLM to process and reason over the acoustic signal itself.

The authors employed a two-stage training process. First, they performed continued pre-training on a massive dataset comprising both speech and general audio events (such as AudioSet (Gemmeke et al., 2017b) and WavCaps (Mei et al., 2023)), which enabled the LLaMA model to learn the association between its linguistic knowledge and the new acoustic tokens from the Whisper encoder. Following this, they performed instruction fine-tuning on a curated dataset consisting of 9M audio-question-answer (AQA) pairs for over 100 diverse audio-related tasks. This dataset included speech recognition, speaker identification, audio captioning, and complex reasoning tasks such as answering questions about sounds within an audio clip (e.g., *“Is the person speaking male or female?”* or *“What is the sound of the animal in the background?”*).

By demonstrating that a single, unified model could achieve strong performance on this wide array of both speech and non-speech audio tasks, LTU-AS provided a powerful proof-of-concept. It established a new paradigm where the rich reasoning capabilities of LLMs could be extended directly to the audio modality, overcoming the information bottleneck of cascaded systems. This work laid the foundation for the emerging field of LALMs and provides the direct architectural precedent for the voice-native model developed in this thesis.

While I have drawn inspiration from the LTU-AS architecture, particularly in the basic idea of combining perception and reasoning modules, my research extends this foundation with specific innovations aimed at enhancing voice understanding and emotional intelligence

in therapeutic conversations. This specialization represents a novel contribution to the field of audio and speech language models, with potential applications extending beyond traditional therapeutic settings to various mental health support contexts.

Chapter 3

METHODOLOGY

3.1 Overview of Research Framework

This chapter details the four-phase research process used to develop and validate the emotionally intelligent voice language model central to this thesis. My approach evolved through an iterative process, beginning with an analysis of end-to-end LALM training, which first established the viability and significant data and resource challenges of implementing a state-of-the-art large audio language model. In response to these findings, I designed and implemented a novel heuristic-based LALM architecture, a clever and resource-efficient approach that conveys acoustic information to a language model through quantized textual descriptors. This heuristic architecture was then subjected to an empirical evaluation against leading audio language models on real-world audio data to identify the most effective model for genuine voice understanding. In the final phase, therapeutic alignment via cross-modal preference optimization, I applied a Direct Preference Optimization (DPO) alignment strategy, using a text-based preference dataset to align the selected model to therapist responses.

3.2 Analysis of End-to-End LALM Training

3.2.1 Architectural Design

The initial idea was to implement a LALM architecture inspired by the dual-component framework of the LTU-AS model for joint audio and speech understanding (Gong et al., 2023b). My planned architecture consisted of two functionally distinct but interconnected modules: a *perception module* to process raw audio and a *reasoning module* to perform comprehension and generate responses.

For the perception module, the plan was to adopt a dual-stream approach. This design was based on the method used in LTU-AS, which effectively captures both linguistic and non-linguistic vocal information. The core of this module was to be the powerful audio encoder from OpenAI’s Whisper model (Radford et al., 2023). The raw audio input would be processed in two parallel streams: 1) the full Whisper model would generate a discrete text transcript, and 2) a Time-and-Layer-Wise Transformer (TLTR) module would process the intermediate representations from the Whisper encoder to generate continuous “soft” audio tokens (Gong et al., 2023a). These continuous tokens are specifically designed to represent the rich paralinguistic cues (e.g., tone, pitch, rhythm) that are lost in a standard text transcript.

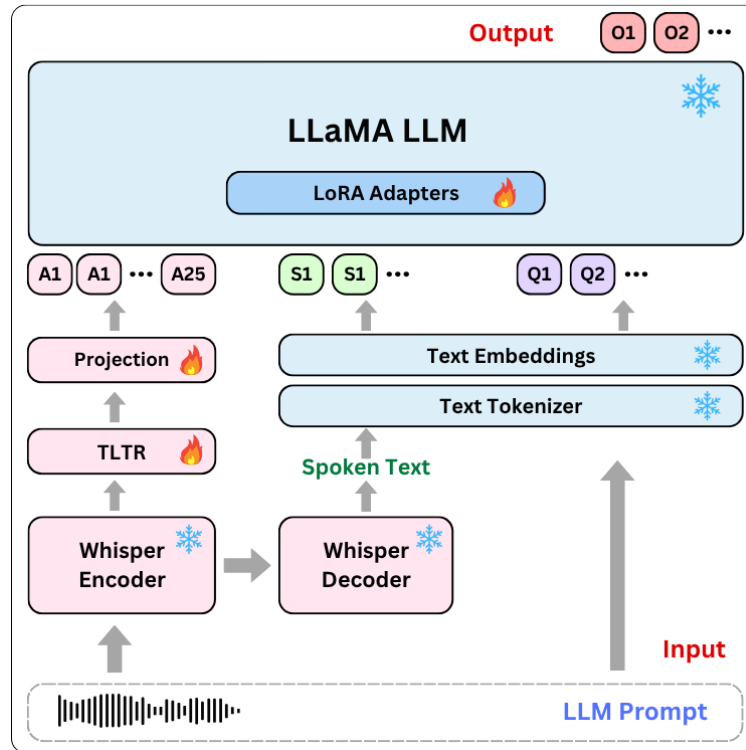


Figure 3.1: LALM Architecture consisting of perception and reasoning modules.

For the reasoning module, my initial approach involved a simpler, zero-shot prompting

strategy. I appended the continuous soft audio tokens directly to the input prompt, along with an explicit instruction for the LLM to interpret them as raw vocal features and perceive the user’s vocal characteristics based on these tokens. This method, however, proved to be ineffective. During preliminary testing, the model suffered from severe hallucinations, failing to ground its responses in the provided acoustic information. We observed that less than 2% of the generated responses were contextually relevant or accurately reflected the perceived vocal cues. This empirical failure demonstrated that the LLM could not inherently understand the semantic meaning of the audio tokens without explicit training.

Hence, I planned to replicate the Vicuna-style instruction tuning process used to train the LLaMA model in the LTU-AS paper (Chiang et al., 2023; Gong et al., 2023b). Accordingly, I selected the `meta-llama/Llama-3.1-8B-Instruct` model for our base reasoning module. The main challenge was to enable this text-centric LLM to understand the composite input from the perception module. As shown in Figure 3.1, my plan was to concatenate the continuous audio tokens ($A1, A2, \dots$) and the decoded speech tokens ($S1, S2, \dots$) with the text prompt tokens ($Q1, Q2, \dots$) into a unified input sequence. To teach the model to interpret this new format while also incorporating it with the necessary domain knowledge, I designed a two-stage fine-tuning process. First, I would fine-tune the base `Llama-3.1-8B-Instruct` model on text-only cultural and mental health datasets (detailed in Section 3.2.2). The goal of this initial stage was to instill mental health therapy knowledge before instruction tuning the model. In the second stage, this domain-aware model would be further fine-tuned on an audio question-answering (AQA) dataset, such as OpenASQA (Gong et al., 2023b), to learn the association between the new audio tokens and its existing language representations. This structured approach represented a direct and theoretically sound path to creating the voice-native language model required by the research problem.

3.2.2 fine-tuning the Reasoning Module on Mental Health Data

My two-stage fine-tuning process began with a text-only supervised fine-tuning (SFT) step. The objective was to provide the reasoning module with a foundation in therapeutic lan-

guage before introducing the additional complexity of audio data. For this task, I used the ShenLab/MentalChat16K dataset from the University of Pennsylvania (Xu et al., 2025b). I selected this dataset primarily because its focus on counseling-style conversations was well-suited for training our model on the appropriate tone and subject matter for therapeutic interactions and mental health contexts. An example from the dataset, illustrating the instruction-following format, is shown below:

Instruction:

You are a helpful mental health counselling assistant; please answer the mental health questions based on the patient’s description. The assistant gives helpful, comprehensive, and appropriate answers to the user’s questions.

Input:

My mom has Alzheimer’s, and I’ve been her primary caregiver for the past few years. I’ve been trying to maintain a routine for her, but my siblings keep calling me for favors or emergencies, which disrupts my schedule and leaves me feeling drained. I’ve tried setting boundaries, but they don’t seem to respect my time or my mom’s needs.

Output:

I’m sorry to hear that your siblings’ demands are causing you such distress. It’s not uncommon for caregivers to face these challenges, and it’s essential to establish clear communication with your family members about your responsibilities and limitations. You might consider setting specific times for responding to their requests or delegating tasks to others. It’s also essential to remember that you’re not alone in this situation and that seeking additional support, such as respite care or community resources, can be beneficial for both you and your mom.

The dataset is composed of two distinct and complementary parts. First, a 6,000-sample corpus derived from real clinical data. This component is sourced from transcripts of behavioral intervention sessions between health coaches and caregivers. Crucially, these interactions are not tailored to a specific therapeutic style like Cognitive Behavioral Therapy (CBT), but rather represent a broader approach of behavioral health coaching. The original dialogues were systematically processed using Mistral-7B-Instruct to distill them into

discrete, high-quality question-answer pairs. This foundation of real-world interactions provides an invaluable source of authentic language, tone, and conversational flow specific to supportive care.

Second, a 10,000-sample synthetic corpus was created to provide the topical breadth and scale necessary for effective fine-tuning. The dataset’s authors employed a two-step generation process using the GPT-3.5 Turbo model and a customized Airoboros framework to create realistic client queries and corresponding counselor responses (Durbin, 2023). This structured approach, covering 33 distinct mental health topics such as relationships, anxiety, depression, intimacy, and family conflict, was designed to mimic the complexity and diversity of real therapeutic interactions. While I acknowledge the inherent limitations of synthetic data, the rigorous methodology used by the dataset’s authors gives this component a high degree of quality and relevance. This made MentalChat16K a powerful and suitable dataset for our SFT.

The fine-tuning was performed on a model we refer to as ‘Carebot’, which is the Llama3.1-8B-Instruct model previously fine-tuned on cultural and mental health datasets as part of the iCare project. The previous datasets utilized for fine-tuning include EmpatheticDialogues (Rashkin et al., 2019), Amod Mental Health (AI, 2023), Psych8k (Liu et al., 2023), ExTES (Ailover, 2024) and the CultureLLM dataset (Li et al., 2024), originally built on the World Values Survey (WVS) (Inglehart et al., 2023). Starting with this already domain-aware model provided a strong foundation for further fine-tuning. I used low-rank adaptation (LoRA), a parameter-efficient fine-tuning (PEFT) technique, to make the process computationally manageable (Hu et al., 2021). In my LoRA technique, I targeted all linear layers for adaptation to allow for widespread adjustments to the model’s learned representations. The key hyperparameters were set as follows: a rank (‘r’) of 64 and a ‘lora_alpha’ of 128. The alpha parameter serves as a scaling factor for the updates, and the 2:1 ratio of alpha-to-rank is a common practice that I adopted to stabilize training. A ‘lora_dropout’ of 0.05 was also applied as a regularization technique to mitigate the risk of overfitting on the data. The complete LoRA-Config and hyper-parameters list is provided in the Appendix B.1.1.

The training itself was implemented using the ‘SFTTrainer’ from the Hugging Face TRL library (von Werra et al., 2020). The training arguments were carefully selected to ensure stable and efficient learning. I used a ‘per_device_train_batch_size’ of 8 with ‘gradient_accumulation_steps’ of 4, resulting in an effective batch size of 32, a standard choice that balances memory usage with gradient stability. The model was trained for 5 epochs with a learning rate of $2e-4$. I enabled mixed-precision training (‘fp16=True’) to reduce memory usage and accelerate training. The maximum sequence length was set to 2048 tokens to accommodate the longer, context-rich dialogues in the dataset. The entire fine-tuning was performed on $1 \times$ RTX A6000 GPU with a total of 48GB VRAM.

To validate the effectiveness of this fine-tuning step, I performed a preliminary evaluation of the SFT-tuned model. I compared its performance against the original **Llama3.1-8B-Instruct** model, the base Carebot model (culturally fine-tuned **Llama3.1-8B-Instrct**), and Google’s state-of-the-art **gemini-2.5-pro** model. The Gemini 2.5 Pro model is chosen as the SOTA model instead of GPT4 models based of its performance and rankings in official benchmarks like LMArena (lma, 2025) and LLM Leaderboard (vel, 2025) as of 05-06-2025. The evaluation was conducted on a held-out test set of 100 samples from the same dataset that were not seen during training. Using a suite of automated metrics, our SFT-tuned model performed comparably to the Gemini 2.5 Pro model and significantly outperformed all other baselines, demonstrating superior knowledge and response relevance for mental health support. The detailed quantitative results of this evaluation are presented in Chapter 4.1. This served as the starting point for the next planned stage of training the model to process and understand acoustic features from raw audio.

3.2.3 Automated Metric-Based Evaluation

I began the metric-based evaluation with response generation. For all four models, I generated responses to each of the 100 inputs in the MentalChat16k dataset. The Gemini 2.5 Pro model was used as an upper benchmark to assess how well the fine-tuned model performed in comparison to the state-of-the-art. I then conducted an automated evaluation by

comparing the generated responses against the ground-truth outputs from the test set. All model responses for this evaluation were generated on a system equipped with $4 \times$ NVIDIA GeForce GTX 1080 Ti GPUs, providing a total of 44 GB of VRAM.

I employed a suite of established metrics designed to measure different aspects of text quality, from lexical overlap and semantic similarity to fluency and diversity. Each metric provides a unique lens for analyzing the models’ outputs.

BLEU: A Precision-Focused N-gram Metric

The Bilingual Evaluation Understudy (BLEU) score is a precision-focused metric that quantifies the overlap of n-grams (contiguous sequences of n words) between the generated text and the reference text (Papineni et al., 2002). It is primarily concerned with how many of the n-grams in the generated response are “correct,” i.e., also present in the ground truth. The score is calculated as a weighted geometric mean of the modified n-gram precisions, multiplied by a brevity penalty to discourage overly short responses.

The formal equation for BLEU is

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (3.1)$$

where p_n is the modified n-gram precision for n-grams of length 1 to N (typically N=4), w_n are uniform weights, and BP is the Brevity Penalty, defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (3.2)$$

Here, c is the length of the candidate response and r is the effective reference length.

The primary function of the BLEU score is to quantify how closely a generated response matches the ground-truth expert response on a lexical level. A higher BLEU score suggests that the model’s output is fluent and uses phrasing that is similar to the reference text. While it is a valuable measure of grammatical and lexical fidelity, its reliance on exact n-gram matches means it does not account for semantic similarity via paraphrasing or synonymy.

ROUGE: Recall-Focused N-gram and Subsequence Metrics

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics that operates as a complement to BLEU, focusing on recall rather than precision (Lin, 2004). It measures how well the generated response captures the information present in the reference text. I used two variants:

ROUGE-N calculates the recall based on n-gram overlap. Its formula is

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Ref}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{Ref}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (3.3)$$

where $\text{Count}_{\text{match}}$ is the number of n-grams co-occurring in both the generated and reference texts.

ROUGE-L measures similarity based on the Longest Common Subsequence (LCS), which does not require the matching words to be contiguous. This makes it more flexible for capturing sentence-level structure. The recall (R_{lcs}), precision (P_{lcs}), and F-measure (F_{lcs}) are calculated as:

$$R_{lcs} = \frac{\text{LCS}(X, Y)}{m} \quad (3.4)$$

$$P_{lcs} = \frac{\text{LCS}(X, Y)}{n} \quad (3.5)$$

$$F_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (3.6)$$

where m and n are the lengths of the reference sentence X and the candidate sentence Y , respectively, and β is a parameter to weight recall.

High ROUGE scores indicate that the model has successfully incorporated the key points and concepts from the ground-truth response. It is particularly important for evaluating whether the core “gist” of the desired answer has been conveyed.

BERTScore: A Semantic Similarity Metric

To evaluate semantic equivalence beyond lexical overlap, I used BERTScore (Zhang et al., 2020). This metric leverages the contextual embeddings from pretrained transformer models

like BERT to compute the similarity between tokens in the candidate and reference sentences. It calculates precision, recall, and an F1 score based on the maximal cosine similarity between token embeddings. The F1 score is formally defined as:

$$F1_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (3.7)$$

where P_{BERT} and R_{BERT} are the precision and recall scores computed from the sums of maximal cosine similarities between the embeddings of the generated and reference sentences.

BERTScore is crucial for this evaluation, as it assesses whether the model has captured the meaning of the reference response, even if different vocabulary is used. In a therapeutic context, where empathy and understanding are key, semantic equivalence is often more important than verbatim repetition.

METEOR and Distinct-n: Advanced Linguistic Quality Metrics

I included two additional metrics to capture more nuanced aspects of linguistic quality.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) improves upon n-gram metrics by incorporating synonymy and stemming and by explicitly modeling word order (Banerjee and Lavie, 2005). It creates an alignment between the candidate and reference sentences and calculates a score based on a weighted F-score and a fragmentation penalty. The score is given by:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Pen}) \quad (3.8)$$

where F_{mean} is a harmonic mean of precision and recall with recall weighted more heavily, and Pen is a penalty based on the number of “chunks” of contiguous matched unigrams. METEOR provides a more holistic measure of translation and paraphrasing quality than BLEU or ROUGE, as it rewards the use of semantically similar words and correct sentence structure.

Distinct-n is an intrinsic metric that measures the lexical diversity of the generated text itself, without comparison to a reference. It is calculated as the ratio of unique n-grams to

the total number of n-grams in the text.

$$\text{Distinct-n} = \frac{\text{Count}(\text{unique n-grams})}{\text{Count}(\text{total n-grams})} \quad (3.9)$$

This metric is important for diagnosing model pathologies like repetition and genericness. A higher Distinct score means the model is creating more varied responses, which is generally viewed as desirable for open-domain chatbots or creative conversation agents because it suggests human-like, engaging dialogue.

The detailed results and discussions are presented in Section 4.1. The complete analysis of these metrics was performed entirely on Google Colab using a T4 GPU.

3.2.4 Audio Instruction Tuning Challenges and Methodological Pivot

With the domain-adapted `Llama-3.1-MentalChat-SFT` model in place, the next planned step was to instruction-tune it on audio data to enable multimodal understanding. This phase, however, revealed a series of significant data acquisition and computational challenges that are common yet often under-documented in the research field.

My initial attempt to integrate the perception and reasoning modules failed immediately. When I provided the projected continuous audio tokens from the TLTR module to our SFT model, it was unable to interpret their meaning, resulting in hallucination and irrelevant outputs. Since my initial attempt to integrate the perception and reasoning modules failed immediately, it confirmed my hypothesis that the LLM required explicit instruction-tuning on a large-scale AQA dataset to learn the semantic relationship between acoustic features and language.

The first major obstacle arose during the download of the OpenASQA dataset, which was used to train the original LTU-AS model. The dataset consists not of audio files but of YouTube video IDs and timestamps. Upon attempting to download the source audio, I discovered that a substantial portion of the links was no longer active due to videos being removed, made private, or placed under new restrictions. This issue, known as “data decay,”

made it impossible to reconstruct the dataset as required for my training plan within the timeframe of this thesis.

As an alternative, I explored the CompAR dataset, another AQA resource used to train the GAMA LALM (Ghosh et al., 2024). While we were able to access a portion of this dataset, my fine-tuning experiments on a 5,000-sample subset yielded poor results. The model continued to generate incoherent responses, suggesting that this dataset was not suitable to effectively teach the model audio feature understanding. Furthermore, I identified a fundamental misalignment: both OpenASQA and CompAR are mostly composed of general-purpose audio (e.g., animal sounds, environmental noises) and are not tailored for the specific nuances of human vocal paralinguistics, which is the core focus of this fine-tuning.

Additionally, the authors of both LTU-AS and GAMA report training their models on large clusters of high-end GPUs (e.g., $4 \times$ Nvidia RTX A6000) for multiple days. This level of computational requirement was beyond the scope of the GPU resources available in our lab, so committing to such a resource-intensive training process was pragmatically infeasible. This constraint, combined with the fact that the pre-trained LTU-AS model was not made publicly available for inference via API, effectively closed the path of direct replication.

These cumulative challenges led to a critical methodological pivot. The initial research plan, while theoretically sound, proved to be impractical. This made me to reformulate my research question to a more resource-efficient alternative: *“How can we provide the acoustic information to the LLM without directly integrating audio feature vectors into the input and avoiding the resource-intensive training?”*. This new question forms the foundation of the heuristic-based LALM architecture.

3.3 Heuristic-Based LALM with Discretized Audio Features

The critical methodological pivot was guided by a new research question: “How can we efficiently convey a user’s vocal acoustics through text to an LLM?” My exploration of this question led to the development of a heuristic-based architecture designed to translate raw acoustic features into discrete, semantic descriptors that an LLM can natively understand.

3.3.1 Initial Techniques and Their Limitations

Before arriving at my final method, I analyzed two initial concepts and their respective drawbacks.

The first concept I explored was a “conditional heuristic model.” This approach involved calculating summary statistics, such as mean magnitude and standard deviation, across the entire sequence of audio tokens generated by the TLTR model. This initial step immediately collapses the time-varying dynamics of speech into a static set of numerical descriptors, discarding crucial temporal information. These statistics were then fed into a series of manually defined conditional statements to map the entire audio segment to a single, predefined descriptive sentence, such as *“high energy speech”*.

While simple to implement, I rejected this approach due to its design being a coarse classification system. By reducing the rich, continuous, and dynamic spectrum of human vocal cues into a single, static label, the method discards the very nuances it is supposed to capture. The inherent temporal variations in speech, the subtle shifts in tone, the pacing, the pauses are all lost in the initial calculation of summary statistics. Consequently, this method cannot represent the complexity of an emotional state as it evolves over time; it can only assign a crude, monolithic classification, thereby losing the significant nuance required for a meaningful therapeutic interaction.

The second concept was a “raw numerical pass-through.” This involved providing the raw statistical scores (e.g., ‘mean_energy: 0.124’) directly in the LLM’s prompt. I concluded this approach was fundamentally weak due to a well-known limitation of LLMs: their inability to effectively reason over raw floating-point numbers. LLMs treat a number like ‘0.124’ as a sequence of discrete tokens (‘0’, ‘.’, ‘1’, ‘2’, ‘4’) rather than as a value on a continuous scale. Without extensive, large-scale fine-tuning, the model has no “grounding” to understand whether ‘0.124’ is high or low, making this method unreliable.

3.3.2 Heuristic Discretization via Semantic Vector Quantization

My final architecture successfully synthesizes the strengths of the previous concepts while mitigating their weaknesses. The chosen approach is a practical application of “Vector Quantization (VQ),” a technique in machine learning where complex, continuous data is represented by a discrete “*codebook*” of symbols (van den Oord et al., 2018). This principle of discretization is also a core component of state-of-the-art speech models like HuBERT, which discretize continuous audio into symbolic units before main processing (Hsu et al., 2021).

The goal of VQ is to map a continuous input vector $\mathbf{x} \in \mathbb{R}^D$ to a discrete index k . This is achieved using a finite codebook, $E = \{\mathbf{e}_i \in \mathbb{R}^D | i = 1, \dots, K\}$, which contains K representative vectors, often called centroids or embeddings. The quantization function, $Q(\mathbf{x})$, finds the index of the codebook vector that is closest to the input vector, typically measured by the squared Euclidean distance. This operation is defined as:

$$k = Q(\mathbf{x}) = \arg \min_i \|\mathbf{x} - \mathbf{e}_i\|^2 \quad (3.10)$$

The output, k , is the discrete index that represents the continuous input vector $\mathbf{x}$. In my heuristic application of this principle, the continuous input vector $\mathbf{x}$ corresponds to the summary statistics calculated from the audio (e.g., a vector of [mean magnitude, standard deviation]). My “codebook” is not a set of learned embedding vectors but rather the predefined numerical boundaries of the semantic tiers. Therefore, the ‘arg min’ operation in Equation 3.10 is implemented not through a distance calculation, but through a series of conditional statements that determine which predefined bin the input vector $\mathbf{x}$ falls into. The resulting index k directly corresponds to a semantic label (e.g., “very low”, “low”, “medium”, “high”, “very high”) for every feature instead of a single, overall semantic label.

The process of my heuristic LALM, which applies this VQ principle, is as follows:

1. Acoustic Feature Summarization: To capture a comprehensive acoustic representation, this process leverages information from both the raw audio waveform and its corresponding high-level feature vectors. First, fundamental acoustic features like pitch,

energy, and mel-frequency cepstral coefficients (MFCCs) are extracted directly from the raw audio waveform, a standard approach for converting complex audio signals into a structured format for analysis (Sharma et al., 2020; Vieting et al., 2021).

Concurrently, the continuous audio feature vectors generated by the perception module, which form a tensor $\mathbf{A} \in \mathbb{R}^{T \times D}$ (where T is time steps and D is the feature dimension), are summarized to capture prosodic dynamics. To create a compact representation of the entire audio clip, I first calculate the L2 norm of the feature vector for each time step, which gives us a sequence of scalar magnitude values $\mathbf{m} = [m_1, m_2, \dots, m_T]$, where $m_t = \|\mathbf{A}_t\|_2$. This technique is often used in signal processing to derive a measure of magnitude while being computationally efficient (Cai et al., 2018). From this sequence, I calculate two key summary statistics that are proxies for prosodic information (Wagner and Watson, 2010):

- The mean magnitude, $\bar{m} = \frac{1}{T} \sum_{t=1}^T m_t$, which serves as a proxy for the overall vocal energy of the utterance.
- The standard deviation (σ_m) of the magnitude sequence $\mathbf{m}$, which serves as a proxy for the prosodic variability or expressiveness of the speech.

This results in a single summary vector, $\mathbf{x} = [\bar{m}, \sigma_m]$, that represents the continuous acoustic properties of the audio clip.

2. Discretization into Semantic Tiers: This continuous summary vector $\mathbf{x}$ is then discretized. This is achieved by mapping each of its components to a predefined, human-readable tier based on empirically derived thresholds. This process, a common practice in classical speech emotion recognition (SER) for handling prosodic features (Schuller and Batliner, 2013), effectively implements the VQ “codebook” as a set of conditional rules. For instance, the vocal energy component, $\bar{m}$, is mapped to one of five semantic tiers:

- ‘0.0 – 0.05’: “Very Low”
- ‘0.05 – 0.1’: “Low”
- ‘0.1 – 0.15’: “Medium”
- ‘0.15 – 0.2’: “High”
- ‘> 0.2’: “Very High”

A similar binning process is applied to the other statistics, such as pitch variability.

3. **Structured Prompt Construction:** Finally, the resulting discrete semantic descriptors are assembled into a structured textual block. This block is appended to the user’s transcribed speech to form the final, context-rich prompt for the LLM. Providing the acoustic analysis in a structured format with clear headings (e.g., ‘Acoustic Analysis:’) ensures that the reasoning module can reliably parse the information and differentiate the user’s words from our analysis of their vocal delivery.

This method excels over the conditional heuristic approach in three key ways. First, it expands information bandwidth: rather than a single label that yields only a few coarse classes, two independent five-tier axes deliver 25 possible acoustic states, giving the language model roughly five times more detail for tailoring tone and empathy. Second, it offers compositional expressiveness: energy and variability are binned separately (e.g., “Energy = High, Variability = Low”), enabling nuanced combinations that the earlier scheme could never capture, and thus supporting more precise dialogue strategies. Third, it ensures future extensibility: each acoustic attribute sits on its own axis, so adding new cues such as pitch range or speech rate requires only setting additional thresholds, whereas the single-label heuristic would demand a complete rewrite of its rule table each time.

3.3.3 *Heuristic Feature Definition*

The first step was to define a set of heuristic features to convey the key paralinguistic characteristics of speech. I selected the heuristic features for this study based on foundational

research in speech processing and computational paralinguistics, which provides a strong justification for their use. The choice of vocal energy, measured via root mean square (RMS), is grounded in its established role as a reliable proxy for perceived loudness and speech intensity (Rabiner and Schafer, 1978). For vocal expressiveness, I adopted the standard deviation of acoustic features, a method validated for its effectiveness in capturing the prosodic contours that differentiate expressive from monotone speech (Eyben et al., 2016). Finally, the inclusion of speech rhythm is justified by its well-documented connection to a speaker’s emotional state and cognitive load, confirming its relevance as a behavioral indicator (Schuller and Batliner, 2013). While this core set of features provides a robust and interpretable starting point, the framework is intentionally designed to be extensible. Future iterations could incorporate a wider array of features to enhance descriptive power, such as spectral characteristics (e.g., Mel-frequency cepstral coefficients), voice quality metrics (e.g., jitter and shimmer), and more granular temporal dynamics. The selection and refinement of such additions would be guided by further literature review and consultation with domain experts to ensure they are both computationally sound and perceptually meaningful.

For this thesis, I selected and implemented three core features:

1. **Vocal Energy:** This feature approximates the perceived loudness or intensity of the speech. For a given raw audio waveform w of length N , we calculate the Root Mean Square (RMS) energy. RMS energy is a standard feature in audio analysis used to estimate the short-term power of a signal, which correlates strongly with perceived loudness (Rabiner and Schafer, 1978):

$$\text{Vocal Energy (RMS)} = \sqrt{\frac{1}{N} \sum_{n=1}^N w[n]^2} \quad (3.11)$$

2. **Vocal Expressiveness:** This feature captures the prosodic variability in a user’s voice, such as changes in pitch and intonation. It is calculated as the standard deviation of the feature values within the distilled audio tokens produced by the TLTR model.

Using the standard deviation of acoustic features as a measure of prosodic contour is a common technique in computational paralinguistics to differentiate between expressive and monotone speech (Eyben et al., 2016)

3. **Speech Rhythm:** This feature represents the perceived pace or cadence of the speech. The rate of speech is a critical cue for interpreting a speaker’s emotional state, such as agitation (fast speech) or sadness (slow speech), and their cognitive load (Schuller and Batliner, 2013). Unlike the other features, which were derived from direct signal processing, the value for this feature was established through the perceptual annotation process detailed in Section 3.3.5

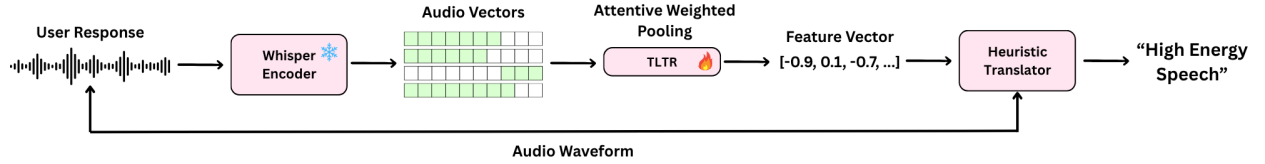


Figure 3.2: Conversion of audio into semantic descriptors using Heuristic Translator

As illustrated in Figure 3.2, the heuristic extraction pipeline begins with the user’s raw audio waveform. This signal is processed in two parallel streams to capture a comprehensive set of acoustic cues. In the primary path, the waveform is fed into a Whisper encoder, which transforms it into a sequence of high-dimensional audio vectors. These vectors are then passed through an attentive weighted pooling layer (TLTR) that condenses the entire sequence into a single, fixed-size feature vector, summarizing the utterance’s key acoustic characteristics. This derived feature vector, along with the original raw audio waveform from the second path, is then supplied to a heuristic translator module. The translator uses both of these inputs to calculate a set of numerical scores corresponding to predefined acoustic heuristics.

This extraction process results in a continuous summary vector $\mathbf{x}$ containing the raw scores for each heuristic. We map each component of the summary vector to one of several

predefined, human-readable semantic tiers (e.g., “Low”, “Medium”, “High”). These resulting semantic descriptors are then assembled into a structured textual block, which is appended to the user’s transcript to form the final, context-rich prompt, as illustrated below.

System:

You are a compassionate, empathetic therapist. Analyze the user’s words and the provided acoustic analysis to form your response.

User Transcript:

”I had a major breakthrough today. I feel amazing.”

Acoustic Analysis:

- Vocal Energy: High
- Pitch Variability: Very High
- Speaking Rythm: Medium

Therapist Response:

3.3.4 Does These Features Contribute to Emotional Intelligence?

My approach to building emotionally aware systems is grounded in established affective science, particularly the Russell Circumplex Model of Affect (Russell, 1980). This model describes emotions using two core dimensions: arousal (the level of energy) and valence (the feeling of pleasantness). Our framework demonstrates a sophisticated understanding of this model by intelligently combining acoustic cues with textual content. The selected features, such as vocal energy and speech rhythm, excel at capturing the arousal dimension of speech. To accurately interpret emotional valence, our system smartly integrates these acoustic signals with the semantic meaning of the words being spoken. This strategy mirrors human perception, where we understand emotion by listening to both what is said and how it is said. For example, a high-energy voice could signal either joyful excitement or anxious fear, and by analyzing the accompanying text, our system can effectively distinguish between the two.

To further enrich this capability and build on our system’s existing emotional intelligence,

we can incorporate additional acoustic features that are more directly tied to valence. This would add another layer of insight, working alongside the textual analysis. For instance, we can introduce measures of voice quality like jitter and shimmer—tiny, rapid variations in pitch and volume—which often correspond to emotional tension. We can also analyze the spectral characteristics of the voice to measure its “brightness” or clarity, using features like the harmonics-to-noise ratio and spectral centroid, which help differentiate a clear, positive tone from a breathy or muted one. Integrating these detailed acoustic markers with our current heuristics and textual analysis would create an even more holistic and human-like understanding of emotion, making our system more robust and less reliant on a single source for valence detection (Picard, 2000).

3.3.5 Empirical Validation of Heuristic Thresholds

Once the final features were decided, it’s crucial to validate the definition of these heuristics. To ensure these thresholds were data-driven and not arbitrary, we conducted a validation study using a 100-sample corpus selected from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset (Livingstone and Russo, 2018). We specifically chose this dataset because it provides recordings of actors portraying a broad and balanced set of discrete emotions (e.g., calm, happy, sad, angry), offering clearer perceptual anchors for our heuristic features than more naturalistic but less emotionally controlled conversational datasets like IEMOCAP. The study was designed to first establish a reliable measure for human perception and then use that measure to derive the final classification boundaries.

The process began with a dual annotation of the corpus. First, to establish a “human ground truth,” each of the 100 audio samples was manually rated on a 1-5 scale for the three heuristic concepts (vocal energy, vocal expressiveness, and speech rhythm). Concurrently, we prompted the Gemini 2.5 Pro model to listen to the same samples and provide a score on a 1-5 scale for each concept, yielding a large-scale set of automated annotations.

To justify using the automated annotations as a reliable measure for human judgment, we then conducted a correlation analysis. We calculated the Pearson correlation coefficient

(r) between the human scale ratings and the scores generated by the Gemini model for each heuristic. Given two paired series $\mathbf{x} = (x_1, \dots, x_n)$ (human ratings) and $\mathbf{y} = (y_1, \dots, y_n)$ (Gemini scores), the coefficient is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3.12)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ are the respective sample means. Equation (3.12) normalizes the covariance of the two variables by the product of their standard deviations, yielding $r \in [-1, 1]$; values near 1 indicate a strong positive linear association, values near -1 a strong negative association, and values near 0 little or no linear relation (Bravais, 1844; Pearson, 1896).

For each heuristic dimension we formed an n -length vector of human ratings and an aligned vector of Gemini outputs. We then applied Equation (3.12) to obtain r , interpreting its magnitude according to conventional thresholds (for example: $|r| < 0.3$ weak, $0.3 \leq |r| < 0.5$ moderate, $|r| \geq 0.5$ strong). The results of this validation are discussed in Section 4.2.

This approach of using an advanced model like Gemini to listen to and rate audio samples is a new and promising direction in research, made possible by the recent progress in artificial intelligence. Today’s large multimodal models are not limited to just understanding text; they can also process and reason about sounds, images, and other types of information. This allows them to perceive complex qualities in audio—such as emotional tone or vocal energy—in a way that is surprisingly similar to human listeners. Several recent studies have shown that these models can be effective at audio-related tasks, like classifying sounds or recognizing emotions in speech (Cappellazzo et al., 2025; Hu et al., 2024; Team, 2025). This growing body of work supports our decision to use Gemini, demonstrating that it is a valid and powerful tool for generating reliable, large-scale annotations in a data-driven way.

3.3.6 Pros & Cons of Heuristic-Based Approach

The decision to use a heuristic-based approach for analyzing paralinguistic features involves a trade-off between interpretability, efficiency, and flexibility. Table 3.1 summarizes the key advantages and disadvantages of this methodology.

Pros	Cons
Very efficient, requires low computational power.	Loses detailed information by using simple tiers.
Transparent and easy to debug when errors occur.	Uses fixed thresholds that may not be optimal for diverse users.
Works with any text-based LLM (and multimodal LLMs).	Can only analyze predefined features, not discover new ones.
Allows direct control over which features are analyzed (explainability).	Requires manual engineering to add or change features.

Table 3.1: A summary of the pros and cons of using a heuristic-based approach for paralinguistic analysis.

One of the primary strengths of a heuristic-based approach is its efficiency and transparency. Because these features are based on simple, well-defined rules, they require very little computational power to calculate, making them ideal for real-time applications. This simplicity also brings a high degree of explainability; when an error occurs, it is straightforward to trace it back to a specific calculation or rule, which makes the system much easier to debug compared to more complex, “black-box” models. Furthermore, this approach provides direct control over exactly which features are being analyzed and allows for seamless integration with any LLM. This makes the approach more scalable for future advancements, as the reasoning module can be replaced with any advanced or more efficient models (like 1.5B parameter models) without needing to modify their underlying architecture.

However, the approach also comes with notable limitations. The use of simple tiers (e.g., “Low,” “Medium,” “High”) to represent continuous acoustic data inherently loses some detailed information, potentially overlooking subtle but meaningful variations in speech. The reliance on fixed thresholds can also be a drawback, as a one-size-fits-all approach may not be optimal for a diverse range of users with different languages and baseline speaking styles. Moreover, the system is constrained to the predefined features it was designed to look for; it cannot independently discover new or unexpected acoustic patterns that might be relevant. Finally, any updates or additions to the feature set require manual engineering, which can be a time-consuming process and may limit the system’s ability to adapt quickly. Despite these cons, the heuristic-based method provides a robust, interpretable, and computationally lean foundation for analyzing speech and voice.

3.4 From Heuristics to Native Voice Understanding

Based on the analysis of pros & cons of this approach, we understood that the heuristic approach is not a substitute for genuine auditory comprehension; it is a tractable, interpretable, and resource-efficient bridge between speech acoustics and text-centric reasoning, thereby allowing targeted threshold adjustment when clinical experts wish to calibrate sensitivity or specificity. While it successfully translates key acoustic features into a semantic format that the LLM can process, it is essential to recognize this as an effective measure rather than a method for true native voice understanding. The conversion of a rich, continuous audio signal into discrete textual descriptors inevitably creates an information bottleneck, sacrificing significant nuance. The reasoning module does not “hear” the audio; it reads a simplified, human-engineered summary.

This architectural choice prevents the model from independently learning the complex, subtle correlations between vocal patterns and semantic meaning that an end-to-end model could potentially discover. Recognizing this limitation was the primary motivation for the next phase of our research: to find audio language models that understand and reason over voice natively and benchmark our heuristic model against them.

The first model selected was **Qwen2-Audio-7B-Instruct**, a powerful LALM that is specifically designed for advanced audio understanding and instruction-following. The Qwen2-Audio technical report demonstrates its state-of-the-art performance, notably outperforming models like Gemini 1.5 Pro on audio-centric instruction-following benchmarks such as AIR-Bench (Yang et al., 2024). Its proficiency across a wide range of audio tasks, including speech recognition and vocal sound classification, established it as a critical benchmark for pure audio understanding (Chu et al., 2024).

The second model selected was **Qwen2.5-Omni-7B**, a more recent and powerful end-to-end multimodal language model designed to perceive diverse modalities including audio, video, and text (Xu et al., 2025a). Its technical report highlights its thinker-talker module architecture and its state-of-the-art performance on voice-specific benchmarks. It surpasses other leading models on tasks like the MMAU benchmark for audio reasoning (Sakshi et al., 2024) and achieves an impressive average score on VocalBench, a benchmark designed to evaluate complex voice-based conversational capabilities (Liu et al., 2025). The Qwen2.5-Omni model, therefore, represents the gold standard for a system that can holistically understand and reason over voice.

While the Qwen models represent the state-of-the-art in audio/text understanding, the ultimate goal of a voice language model is a voice-to-voice (V2V) architecture—one that not only understands voice but also responds directly with synthesized voice. During our research, a model embodying this principle was recently introduced: Voila, a voice-language foundation model (Shi et al., 2025). Voila’s architecture represents a significant advancement by adopting an end-to-end, hierarchical, multi-scale, transformer-based approach that integrates a voice-language LLM backbone with a dedicated audio transformer.

Rather than bypassing text generation, Voila explicitly models semantic and acoustic tokens separately, enabling simultaneous high-level semantic understanding and low-level acoustic detail synthesis. Its voice tokenizer, built with residual vector quantization, extracts discrete semantic and acoustic tokens from raw audio, which are added to the text vocabulary to facilitate cross-modal training and generation.

Voila supports full-duplex conversational interactions with extremely low latency (around 195 milliseconds), surpassing human conversational response times. It also enables rich vocal nuance preservation, such as tone, rhythm, and emotion, and allows easy speaker customization through text prompts defining speaker attributes or from short audio samples. Beyond spoken dialogue, it unifies various voice-based tasks under a single model architecture, including ASR, TTS, and multilingual speech translation, all trained via a chat-style next-token prediction objective. Voila’s design emphasizes simultaneous listening, reasoning, and speaking in a continuous voice-to-voice flow, emulating natural human interactions rather than a traditional sequential pipeline. Its architecture and capabilities make Voila a significant step towards autonomous voice agents that proactively listen, reason, and respond in real-time with expressive, persona-aware speech.

Therefore, we have selected the `Qwen2-Audio-7B-Instruct`, `Qwen2.5-Omni-7B`, and `maitrix-org/Voila-chat` models for further comparison and benchmarking on voice understanding and responding. All of these models are accessed from HuggingFace.

3.5 Comparative Benchmarking and Model Selection

To contextualize the performance of our heuristic LALM, we conducted a comparative benchmark against the baseline ASR+LLM architecture and the three leading end-to-end models.

3.5.1 *Benchmark Dataset Preparation*

A critical requirement for our benchmark was a dataset that contains genuine, spontaneous human speech and reflects real psychological states. This requirement immediately disqualifies many common academic datasets, such as IEMOCAP (Busso et al., 2008) or RAVDESS (Livingstone and Russo, 2018), which consist of speech from actors portraying discrete emotions. While valuable for some tasks like emotion recognition, such acted data lacks the nuance, subtlety, and non-prototypical expressions found in authentic conversations about mental health.

The field of computational mental health analysis suffers from a significant scarcity of high-quality, naturalistic audio data, largely due to the immense ethical, privacy, and logistical challenges involved in collecting and sharing such sensitive information. The Distress Analysis Interview Corpus (DAIC-WOZ) (Gratch et al., 2014) is one of the very few publicly available resources that overcomes these barriers. It contains audio recordings of human participants in real clinical interviews designed to screen for psychological distress, providing a rare and valuable source of authentic conversational data in a therapy context. While DAIC-WOZ is itself focused on a specific diagnostic task (depression screening), its use of unscripted, spontaneous speech from individuals discussing their well-being makes it the most suitable choice from a very limited set of options for our benchmark.

For our study, only the raw audio from the participant side of the interviews was utilized; the original depression classification labels were not relevant to our task of evaluating general conversational and emotional awareness. The dataset was meticulously segmented at utterance boundaries using transcription timestamps to yield 30–50 second test clips, facilitating consistent and granular evaluation. This corpus of naturalistic speech segments served as the input for all models in our benchmark. All of the dataset processing was performed in 2 days on $4 \times$ NVIDIA GeForce GTX 1080 Ti GPUs, providing a total of 44 GB of VRAM.

3.5.2 Models Under Evaluation

The comparative analysis included five distinct architectures:

1. **Baseline ASR + SFT LLM:** The standard architecture is where speech is first transcribed to text by the Whisper model, and the resulting text is fed to our fine-tuned `Llama-3.1-MentalChat-SFT` model.
2. **Heuristic LALM:** Our novel architecture, which provides the `Llama-3.1-MentalChat-SFT` model with both the decoded speech and the text-based semantic descriptors of the audio.

3. **Qwen2-Audio-7B-Instruct:** Provided with only the raw audio clip from the DAIC-WOZ test set. It was tasked with generating a text-based response by internally processing the audio.
4. **Qwen2.5-Omni-7B:** Received only the raw audio clip as input. The benchmark was designed to test its native ability to reason over vocal information and generate a text response.
5. **Voila-Chat:** Received only the raw audio clip as input and generated a text response.

Even though the Qwen2.5-Omni and Voila could generate voice responses, we disabled voice output to mitigate bias and have fair benchmarking of all the five models.

3.5.3 Evaluation Framework: *LLM-as-a-Judge*

For each audio clip in our prepared DAIC-WOZ test set (345 audio samples), a response was generated from all five models using the same prompt (see Appendix B.2.1 for the specific prompts used for response generation). Given the open-ended nature of the task, a standard ground-truth reference for each response does not exist. Therefore, I adopted the LLM-as-a-judge paradigm for evaluation. This approach, where a powerful LLM is prompted to score the outputs of other generative models, has been shown to be a scalable and cost-effective method for assessing model quality. Seminal work in this area demonstrated that state-of-the-art models like Gemini 2.5 Pro and GPT-4, when used as judges, can achieve high agreement with human evaluators (over 80%), making this a reliable and common method for assessing generative models, particularly in open-ended conversational tasks (Zheng et al., 2023)

For this role, I selected the `gemini-2.5-pro-preview` model, which was considered the state-of-the-art in large-scale language understanding at the time of my experiments (05-2025). The Gemini 2.5 Pro model was prompted to rate each response on a set of criteria adapted from the ESHRO framework, an innovative evaluation framework for AI-driven

mental health assistants (Sherwani et al., 2025). These criteria included empathy, relevance, and the apparent awareness of the user’s underlying emotional state; the full details of this evaluation prompt can be found in Appendix B.2.1. Gemini 2.5 Pro rated responses across four critical dimensions: Empathy, Safety, Helpfulness, and Overall Quality on a 10-point scale, accompanied by detailed qualitative justifications. This evaluation framework allowed for a nuanced, qualitative comparison of the different architectural approaches. The detailed quantitative results and analysis of this benchmark are presented in Chapter 4.3.

3.5.4 Performance Verification on VoiLa-Bench

During the initial benchmarking on the DAIC-WOZ dataset, I observed an unexpected result: the VoiLa model, despite its advanced voice-to-voice architecture, performed less effectively than our heuristic LALM and the two Qwen models. Similarly, the Qwen2-Audio model, despite its native audio and text understanding ability, performed poorly in the DAIC-WOZ evaluations. To investigate this surprising outcome and to more directly compare the state-of-the-art native voice understanding models, I conducted a second, more targeted benchmark using the evaluation set provided by the authors of VoiLa.

The Voila-Benchmark, introduced in the official Voila paper (Shi et al., 2025), is a comprehensive audio-language evaluation suite comprising approximately 1,400 audio-text pairs. The questions and reference answers are sourced from five diverse and widely used LLM evaluation datasets: MMLU (Hendrycks et al., 2021a), MATH (Hendrycks et al., 2021b), OpenAI HumanEval (Chen et al., 2021), NQ-Open (Kwiatkowski et al., 2019), and GSM8K (Cobbe et al., 2021). This composition ensures broad domain coverage and provides a robust framework for evaluating the models’ reasoning and comprehension abilities directly from spoken input. As per the Voila authors, the responses of all the models are evaluated by GPT-4o model as a judge. We ran the VoiLa-Bench evaluation on our top three models: Qwen2.5-Omni, Qwen2-Audio-7B-Instruct, and Voila-Chat itself. Our heuristic LALM was excluded from this specific test, as the results from our DAIC-WOZ evaluation, discussed in Section 4.3 had already shown it to be outperformed by the Qwen2.5-Omni model.

This subsequent test focused exclusively on isolating and comparing the capabilities of native voice-understanding architectures. This allowed me to directly verify the performance claims from the VoiLa paper and to understand the relative strengths of the three leading models in a controlled setting. The detailed results and analysis of this benchmarking are presented in Section 4.3.2.

3.5.5 Validation of Voice Language Model Based on VocalBench

To further strengthen my selection of the best-performing model through data-driven and robust approaches, I analyzed the results from a recent benchmark named VocalBench (Liu et al., 2025). This approach was designed to systematically evaluate the performance of the `Qwen2.5-omni-7B` model and assess its superior, well-rounded performance across the most critical attributes for an advanced conversational AI: voice understanding, semantic reasoning, and emotional intelligence.

The methodology involved a comparative analysis of the data presented in Table 2 of the VocalBench study. I began by identifying the core competencies required for a state-of-the-art voice agent, grouping the various VocalBench metrics into three essential pillars. The first, Semantic and Reasoning Capability, was assessed via the *Reasoning Score* and *Knowledge Accuracy* to measure the model’s core intelligence. The second, Fundamental Voice Understanding, was evaluated using the *Clarity* (WER) metric, which quantifies the model’s ability to accurately perceive spoken language. The third pillar, Conversational and Emotional Intelligence, was judged by the *Multi-Round Score* and the *Emotion Score*, which together assess the model’s ability to engage in dynamic, context-aware, and emotionally resonant dialogue. The detailed analysis is provided in Section 4.3.3.

By employing this structured, multi-dimensional evaluation methodology, I can confidently assert that the chosen model is not merely a strong performer on one or two metrics but is demonstrably the most advanced and well-rounded model for complex, real-world voice interaction, excelling in voice understanding, reasoning, and inherent emotional intelligence.

3.6 User-Based Study for Model Selection

While the automated benchmarks provided a quantitative comparison of the models’ technical capabilities, a user-based study was essential to evaluate their real-world effectiveness, perceived empathy, and safety. This section details the methodology for the user study I conducted to compare the final candidate model.

3.6.1 *Ethical Considerations and IRB Approval*

Given that the study involved human subjects interacting with AI-generated responses in the context of mental health, ensuring the ethical treatment and confidentiality of participants was paramount. We obtained formal approval from the Institutional Review Board (IRB) of the University of Washington Human Subjects Division (HSD) on 05-29-2024 (IRB ID: STUDY00020524) to conduct this research. This certification ensured the study complied with all necessary regulations, protected the rights and safety of the participants, and included a clear informed consent process that outlined the purpose of the research and the potential risks and benefits involved.

3.6.2 *Study Design and Procedure*

The study was designed as a comparative evaluation and administered via an online survey. I employed a screening questionnaire at the beginning of the survey that verified that participants were between the ages of 18 and 65, had experience using web applications, and did not have any hearing conditions that would impair their perception of the audio clips. I also collected demographic information regarding their history with mental health services and their field of work or study to better understand the participant cohort. Then, participants were presented with a series of 10 audio clips with humans speaking in different contexts. For each audio clip, they were shown two distinct responses, anonymized and labeled simply as Model A and Model B.

The core of the study involved participants listening to an audio clip and then reading

the two corresponding model responses. Their task was to evaluate and compare the two responses based on a structured set of questions. This A/B testing format was chosen to elicit direct comparative feedback and to minimize bias, allowing us to directly measure user preference between the two models being compared in each trial. To help prevent participants from developing a bias for one model over the other (e.g., always preferring the first response), I swapped which model was shown as ‘Model A’ and ‘Model B’ for every participant halfway through the series of trials. While this measure helps reduce order bias, a more robust approach for future studies would be to fully randomize the presentation order for each trial, ensuring that neither model gains an advantage simply by its position.

3.6.3 Evaluation Method

The survey was designed to capture both quantitative ratings and qualitative judgments. For each pair of responses, participants were asked to answer the following:

1. **Overall Preference:** A direct question asking which response they preferred overall (Model A, Model B, Both, or Neither).
2. **Overall Quality Rating:** A scale of 1-10 rating for both Model A and Model B individually.
3. **Qualitative Metrics:** A scale of 1-5 (Very Poor to Excellent) to rate both models on four criteria adapted from the ESHRO framework (Sherwani et al., 2025): Empathy, Supportiveness, Relevance, and Safety.
4. **Perceptual Accuracy:** Two multiple-choice questions asking which model more accurately understood the speaker’s key points and which better reflected their underlying tone of voice.
5. **Safety Check:** A question to flag if either response was perceived as unsafe, dismissive, or potentially harmful.

6. **Open-Ended Feedback:** A final optional text box for any additional comments.

This holistic evaluation provided a comprehensive view of each model’s performance from the user’s perspective. The detailed results and analysis of this user study are presented in Chapter 4.3.4.

3.7 Alignment of Voice Language Model for Therapy

The comparative benchmarks and the user study identified the Qwen2.5-Omni model as the most capable and the best-performing model for native voice understanding. However, raw performance on comprehension tasks does not inherently equate to emotional intelligence. The final and most crucial phase of my research was to align this powerful model, refining its conversational style to be more empathetic, supportive, and therapeutically effective. This section details the challenges I faced and the alignment strategy I implemented to achieve this goal.

3.7.1 Limitations For Post-training

The primary challenge in fine-tuning a model for mental health applications is the scarcity of high-quality, specialized audio data. For this purpose, I selected the same MentalChat6K dataset (6,000 samples out of 16,000) consisting of real therapeutic interview data and covering 33 distinct mental health topics (Xu et al., 2025b). Its basis in real human conversations provided the authenticity necessary for our alignment goal. I purposefully excluded the synthetically generated 10,000 samples of data to avoid alignment on GPT-generated data as the most recent Qwen2.5-Omni model has proved to be better than the GPT 3.5 model.

My initial consideration was supervised fine-tuning (SFT). However, I concluded that the dataset, while high-quality, was likely insufficient to effectively fine-tune a model of the scale of Qwen2.5-Omni-7B without a significant risk of overfitting. This conclusion was reinforced by recent research into the alignment of state-of-the-art audio models, which demonstrated that Group Relative Policy Optimization (GRPO), a reinforcement learning technique, sig-

nificantly outperformed SFT on the Qwen2-Audio-7B model, even with a relatively small dataset of only 38k samples (Li et al., 2025a). This provided strong evidence that preference optimization techniques could offer a more effective and data-efficient path to alignment. This inspired me to pivot my strategy towards preference-based methods, leading to the core research question: *“Can text-based Direct Preference Optimization (DPO) effectively align a multimodal model like Qwen2.5-Omni for a specialized downstream task?”*. Answering this question is significant, as a positive result would provide a viable framework for aligning voice language models in low-data domains like mental health.

3.7.2 Dataset Preparation for Alignment

A prerequisite for DPO is a dataset structured in a preference format, typically containing triplets of a prompt, a preferred (chosen) response, and a dispreferred (rejected) response. As the MentalChat6K dataset exists in a standard instruction-input-output format, I had to first transform it into a suitable structure for DPO. This transformation was achieved through the following multi-step process:

1. **Test Set Isolation:** To ensure a robust and unbiased evaluation of my final aligned model, I first partitioned the MentalChat6K dataset. A random subset of 500 samples was held out as a dedicated test set, which was not used in any part of the training or alignment process.
2. **Generation of “Rejected” Responses:** The core of my data transformation involved generating the dispreferred, or “rejected,” responses. For every prompt in the remaining training portion of the dataset, I used the base, unaligned Qwen2.5-Omni-7B model to generate a response. These model-generated outputs served as the “rejected” component of the preference pairs. These model-generated outputs were designated as “rejected” because the corresponding expert-written answers in the dataset are fundamentally preferred, creating a clear preference pair for the optimization algorithm. This strategy provides a direct training signal, explicitly teaching the model to shift its

internal policy away from its default conversational style and towards one that more closely emulates the nuances and quality of the expert ground truth.

3. **Assembly of the DPO Dataset:** With the rejected responses generated, I assembled the final dataset for DPO. Each data point in this new dataset consisted of a triplet:

- **Prompt (x):** The original user prompt from the MentalChat6K dataset.
- **Chosen Response (y_w):** The “winner” response, which was the corresponding expert-written, ground-truth output from the MentalChat6K dataset.
- **Rejected Response (y_l):** The “loser” response, which was the answer generated by the base Qwen2.5-Omni-7B model for the given prompt.

This process yielded a preference dataset, enabling me to directly apply the DPO algorithm to align the base model’s outputs more closely with those of a human therapist.

3.7.3 *Alignment via Direct Preference Optimization*

I selected Direct Preference Optimization (DPO) as my alignment method (Rafailov et al., 2023). DPO offers a more direct and often more stable approach than traditional Reinforcement Learning from Human Feedback (RLHF), as it reframes alignment as a simple classification problem on preference data, eliminating the need to train a separate reward model.

DPO optimizes a policy model π_θ to increase the likelihood of chosen responses over rejected ones. This is achieved using a static preference dataset, $\mathcal{D} = \{(x, y_w, y_l)\}_i^N$ where, in the context of this implementation, x represents the “prompt” from the MentalChat6K dataset, y_w is the “chosen” response (the ground-truth expert answer), and y_l is the “rejected” response (the output generated by the base model). The DPO loss function is then defined as

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (3.13)$$

Here, π_{ref} is a static, frozen copy of the initial model policy, β (beta) is a hyperparameter controlling the strength of the preference modeling (typically set to 0.1), and σ (sigma) is the logistic function. In essence, this objective function directly encourages the policy model π_{θ} to increase the relative log-probability of the “chosen” responses while decreasing that of the “rejected” responses. The regularization against the reference policy π_{ref} is crucial, as it prevents the model from deviating too far from its original knowledge base, which helps mitigate catastrophic forgetting.

The choice of DPO over a method like GRPO for this research was driven by its significant advantages in implementation simplicity and computational efficiency. GRPO requires a multi-stage, online training loop involving sampling, reward model inference, and policy gradient calculations, making it more computationally expensive and complex to implement. DPO, by contrast, operates on a static, precompiled preference dataset and requires only a single modification to the loss function, making it as straightforward to implement as standard supervised fine-tuning. This simplicity makes DPO a more practical and accessible choice for aligning models, particularly in research settings where computational resources may be constrained.

3.7.4 Implementation and Evaluation of DPO Alignment

With the preference dataset constructed, I proceeded with implementing text-based DPO on Qwen2.5-Omni models and evaluation.

1. **DPO Training:** I performed DPO training on the Qwen2.5-Omni-7B model using the constructed dataset. The training was implemented using the efficient MS-Swift (ModelScope-Swift) framework (Zhao et al., 2024), which is highly optimized for large model training. The resulting DPO-aligned model was then saved and pushed to the Hugging Face Hub for accessibility and replicability.
2. **Comparative Evaluation:** To measure the effectiveness of the alignment, I generated responses for the 500-sample test set using both the base Qwen2.5-Omni model and

the new DPO-aligned model.

3. **Automated Metrics:** I conducted an automated evaluation comparing both sets of generated responses against the ground-truth expert responses from the test set. I used several standard text-generation metrics, including BERTScore (for semantic similarity), BLEU (for n-gram precision), and METEOR (for a more comprehensive alignment of unigrams).

The results of this analysis, detailed in Chapter 4.4, demonstrated that the DPO-aligned model significantly outperformed the base model across all automated metrics. This confirmed that this text-based DPO approach was successful in aligning the multimodal model to produce responses measurably closer to those of a human therapist, thereby enhancing its emotional intelligence. As a final validation step, we plan to deploy this DPO-aligned model on a Hugging Face Space to gather qualitative user feedback on its perceived therapeutic effectiveness.

3.7.5 Human Expert Evaluation of DPO Model

For this analysis, voice-based responses were generated for 11 carefully selected audio clips, including some produced using the latest Eleven Labs V3 model to simulate expressive speech. Two sets of voice responses were generated for each audio clip: one from the base model and another from the DPO model. Both sets were generated natively by the models without relying on an explicit TTS model, showcasing their end-to-end voice language modeling capabilities.

These paired responses were deployed through a Hugging Face space (Raviteja, 2025) embedded within a Google Form, where human evaluators, specifically domain experts in therapy, unlike the general users in the initial study, were invited to assess the quality of each model’s outputs. The evaluators were asked to rate each model across a series of detailed questions designed to capture different dimensions of performance. For the base model, participants rated overall satisfaction across all samples they heard. They then provided

more specific ratings on the base model’s ability to understand both vocal and contextual information, how well it perceived the emotions expressed, and its demonstration of empathy. Quality of voice naturalness was also rated. To further evaluate emotional support, evaluators assessed the base model on several key affective qualities, including empathy, supportiveness, helpfulness, and relevance to the conversational context.

The exact same set of questions was asked about the DPO model, allowing for a direct side-by-side comparison using consistent criteria. This helped in understanding nuanced differences in emotional intelligence, voice quality, and contextual comprehension between the two models.

Specifically, the questionnaire included items such as

- Rate the base model across all the samples you have heard overall.
- Rate the base model based on voice and context understanding.
- Rate the base model based on how well it perceived emotions.
- Rate the base model on naturalness of voice in its responses overall.
- Please rate the base model on the following qualities: empathy, safety, supportiveness/helpfulness, relevance. (Sherwani et al., 2025)

The exact same set of questions was used for the DPO model to enable a direct comparison between the two models on all these dimensions.

This human evaluation is absolutely essential for addressing a critical question in this research: whether the text-based DPO process genuinely improves a model’s ability to understand voice and respond with greater empathy compared to the base model. While automated metrics and internal model scores offer some insights, they often cannot fully capture the subtle nuances of emotional alignment, contextual understanding, and compassionate response that are crucial in real-world interactions. By involving human experts in therapy to

directly assess and compare paired voice responses, we gain invaluable insights into how the DPO process truly impacts the perceived emotional intelligence and personalization of these models. Demonstrating improved alignment through this human evaluation would provide strong support for using text-based DPO techniques to further optimize models in domains where voice data might be limited, but high-quality emotional alignment is paramount. This makes the human expert evaluation essential for validating the practical benefits and broader applicability of the DPO approach in challenging, data-constrained settings. The detailed results and discussion of this evaluation are presented in Section 4.4.2.

Chapter 4

EXPERIMENTS & RESULTS

4.1 Supervised Fine-tuning of Reasoning Module in LALM

As discussed in the Section 3.2.3, this analysis focuses on comparing the performance of our fine-tuned model, `Llama-3.1-MentalChat-SFT`, against the base model (`Llama-3.1-8B-Instruct`), a culturally tuned variant, and a powerful, proprietary baseline (`gemini-2.5-pro`). The objective is to interpret the performance differences and to understand the impact of our domain-specific SFT strategy.

4.1.1 Analysis of Semantic Equivalence (*BERTScore*)

The analysis of semantic equivalence using *BERTScore* is fundamental to evaluating the efficacy of the SFT strategy, as this metric assesses the model’s ability to capture the intended meaning of the reference responses. The results, presented in Table 4.1, show a nuanced distribution of performance across the dimensions of precision, recall, and their harmonic mean (F1-score).

Table 4.1: Mean BERT scores for each model. Best scores are in bold.

Model	BERT-Precision	BERT-Recall	BERT-F1
Llama-3.1-MentalChat-SFT	0.8649	0.8486	0.8565
gemini-2.5-pro	0.8444	0.8487	0.8465
Llama-3.1-cultural	0.8552	0.8199	0.8370
Llama-3.1-8B-Instruct	0.8666	0.8343	0.8501

A detailed examination of these results reveals several key insights. First, the base `Llama-3.1-8B-Instruct` model achieves the highest precision score (0.8666). This indicates that the general instruction-tuning of the base model already optimizes for generating responses that are highly factual and contain less semantically irrelevant tokens. The `Llama-3.1-MentalChat-SFT` model’s precision is a very close and the second best at 0.8649. The fact that our domain-specific fine-tuning resulted in only a marginal decrease in precision is a positive outcome, as it suggests the SFT process did not degrade the model’s core ability to produce accurate and on-topic responses.

Second, the analysis of recall provides the strongest evidence for the benefit of my SFT strategy. The `gemini-2.5-pro` model, likely due to its immense scale, achieves the highest recall (0.8487). However, the `Llama-3.1-MentalChat-SFT` model’s recall of 0.8486 is statistically identical, and it represents a substantial improvement over its base model’s score of 0.8343. From this result, I infer that the SFT process specifically taught our model to be more comprehensive and to better capture the full semantic content of the expert-written reference texts. It learned not just to be correct but to be thorough in its therapeutic responses.

Finally, and most importantly, the BERT-F1 score, which provides a balanced measure of precision and recall, demonstrates the overall superiority of the fine-tuned model. With a score of 0.8565, the `Llama-3.1-MentalChat-SFT` model outperforms all other models, including the much larger Gemini. While other models excel at either precision or recall, the SFT model achieves the most effective optimization of the precision-recall trade-off. This leading F1 score provides strong quantitative evidence that the SFT procedure was successful in creating a model that is not only highly accurate but also comprehensive, resulting in the highest overall semantic alignment with the desired therapeutic dialogue style.

4.1.2 Analysis of Lexical Precision (BLEU Score)

The BLEU score is used to evaluate the lexical precision of the generated text by measuring the overlap of n-grams with the reference text. The results of this analysis are presented in

Table 4.2.

Table 4.2: Mean BLEU scores (n=1-4). Best scores are in bold.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Llama-3.1-MentalChat-SFT	0.1736	0.0829	0.0444	0.0235
gemini-2.5-pro	0.3291	0.1577	0.0837	0.0464
Llama-3.1-cultural	0.0506	0.0203	0.0098	0.0044
Llama-3.1-8B-Instruct	0.1367	0.0721	0.0426	0.0243

The analysis of these results begins with acknowledging the performance of the **gemini-2.5-pro** model, which achieves the highest scores across all n-gram levels. This outcome is expected, given the scale and massive parameter count. In this context, the fine-tuned model demonstrates a clear and consistent improvement over the other models, particularly on lower-order n-grams. The SFT model’s BLEU-1 score of 0.1736 is substantially higher than the base and cultural models. This indicates a marked improvement in unigram precision, from which I infer that the SFT process successfully taught the model to use more relevant, domain-specific terminology found in the ground-truth therapeutic dialogues.

The same results are observed with bigram precision as well. The SFT model’s BLEU-2 score of 0.0829 is significantly higher than the other models, suggesting that the model has learned not just individual keywords but also common and appropriate two-word phrases and collocations characteristic of a therapeutic conversation (e.g., “it sounds,” “how does,” “feeling overwhelmed”).

For higher-order n-grams (BLEU-3 and BLEU-4), which measure the precision of longer, exact phrases, the SFT model continues to outperform the base and cultural models on BLEU-3 (0.0444 vs 0.0426). The base model scores marginally higher on BLEU-4, though the difference is minimal (0.0243 vs 0.0235) and likely not statistically significant. This indicates that our SFT model has learned to generate more precise language at nearly all

levels of granularity.

4.1.3 Analysis of Lexical Recall (ROUGE Score)

The ROUGE score is used to evaluate the lexical recall of the generated text, measuring the extent to which it captures the content of the reference text. The results of this analysis are presented in Table 4.3.

Table 4.3: Mean ROUGE scores for each model. Best scores are in bold.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Llama-3.1-MentalChat-SFT	0.2911	0.0607	0.1327
gemini-2.5-pro	0.3801	0.0740	0.1571
Llama-3.1-cultural	0.1722	0.0241	0.0918
Llama-3.1-8B-Instruct	0.2760	0.0612	0.1400

Consistent with the BLEU score results, the `gemini-2.5-pro` model achieves the highest scores across all ROUGE metrics. Notably, the `Llama-3.1-MentalChat-SFT` model shows an improvement in the ROUGE-1 score over the base model (0.2911 vs. 0.2760). This indicates that the SFT process improved the model’s ability to recall and utilize key individual terms and concepts present in the ground-truth responses. The model has become more effective at identifying and including the core vocabulary of therapeutic dialogue.

However, the SFT model performs on par with, but marginally lower than, the base model in ROUGE-2 (0.0607 vs. 0.0612) and ROUGE-L (0.1327 vs. 0.1400). This subtle result is highly significant. I infer that while the fine-tuning enhanced the model’s ability to use the correct individual words (improving ROUGE-1), it did not constrain it to use the exact same phrases or sentence structures as the reference text. The model is demonstrating a capacity for generative flexibility, arranging the key concepts it has learned to recall into

novel phrasal and sentence constructions. This behavior is indicative of a deeper semantic understanding rather than rote memorization of structure.

In contrast, the `Llama-3.1-cultural` model scores the least scores across all ROUGE metrics. I hypothesize that this performance is not indicative of a model failure, but rather a direct consequence of its specialized tuning. The cultural model is presumably optimized to generate responses tailored to diverse cultural contexts, introducing inherent variability into its output. This design objective is fundamentally misaligned with the evaluation task of matching a single, specific ground-truth style present in the MentalChat dataset, which naturally leads to lower lexical overlap and thus, lower ROUGE scores.

4.1.4 Analysis of Linguistic Quality Metrics (METEOR & Distinct)

The final set of automated metrics, METEOR and Distinct-n, evaluates more sophisticated aspects of the generated language, including the quality of paraphrasing and the degree of lexical diversity. The METEOR score provides a more advanced measure of alignment than n-gram metrics by accounting for synonymy, stemming, and word order. The results of this evaluation are presented in Table 4.4.

Table 4.4: Mean METEOR and Distinct scores for each model.

Model	METEOR	Distinct-1	Distinct-2
Llama-3.1-MentalChat-SFT	0.1799	0.7894	0.9960
gemini-2.5-pro	0.2629	0.5117	0.8982
Llama-3.1-cultural	0.0892	0.8512	0.9963
Llama-3.1-8B-Instruct	0.1517	0.6355	0.9256

The `gemini-2.5-pro` model, as expected, achieves the highest METEOR score (0.2629), which suggests it is highly proficient at generating fluent and structurally sound paraphrases

that align well with the reference text. The SFT model’s score of 0.1799 represents a substantial improvement over the base model’s score of 0.1517, making it the best-performing model in the 8B parameter class. From this, I infer that the fine-tuning process did more than just teach the model to use correct keywords (as indicated by ROUGE-1); it also improved the model’s fundamental ability to generate linguistically coherent and semantically equivalent paraphrases.

On the other hand, the Distinct-n metrics are intrinsic measures used to evaluate the lexical diversity of a model’s output, with higher scores indicating less repetitive and more varied language. The results in Table 4.4 reveal that the cultural model achieves the highest scores for both Distinct-1 (0.8512) and Distinct-2 (0.9963). I hypothesize that this is a direct consequence of its fine-tuning process, as a model tuned to address diverse cultural contexts would naturally be optimized for generating a wider variety of expressions. However, as established by its low scores on all reference-based metrics, this diversity comes at the expense of relevance to the specific task.

The key insight for my research is the performance of the SFT model. It achieves a Distinct-2 score of 0.9960, which is statistically comparable to the top-performing cultural model. It demonstrates that our SFT model achieves near-maximal lexical diversity while also achieving the highest semantic alignment score (BERT-F1). The fine-tuning process successfully guided the model toward the target therapeutic style without forcing it into a repetitive or generic mode of expression.

This contrasts with the **gemini-2.5-pro** model, which, despite its scale, has a notably lower Distinct-2 score (0.8982). This suggests that very large models may sometimes converge on safer, more canonical, but less varied phrasing. The ability of the SFT model to be both highly relevant and highly diverse is, therefore, a key success of my approach.

4.1.5 Conclusion and Final Model Selection

To synthesize these findings, Figure 4.1 presents a direct comparison of the three Llama 3.1 variants, aggregating the average scores for each metric category. This visualization makes

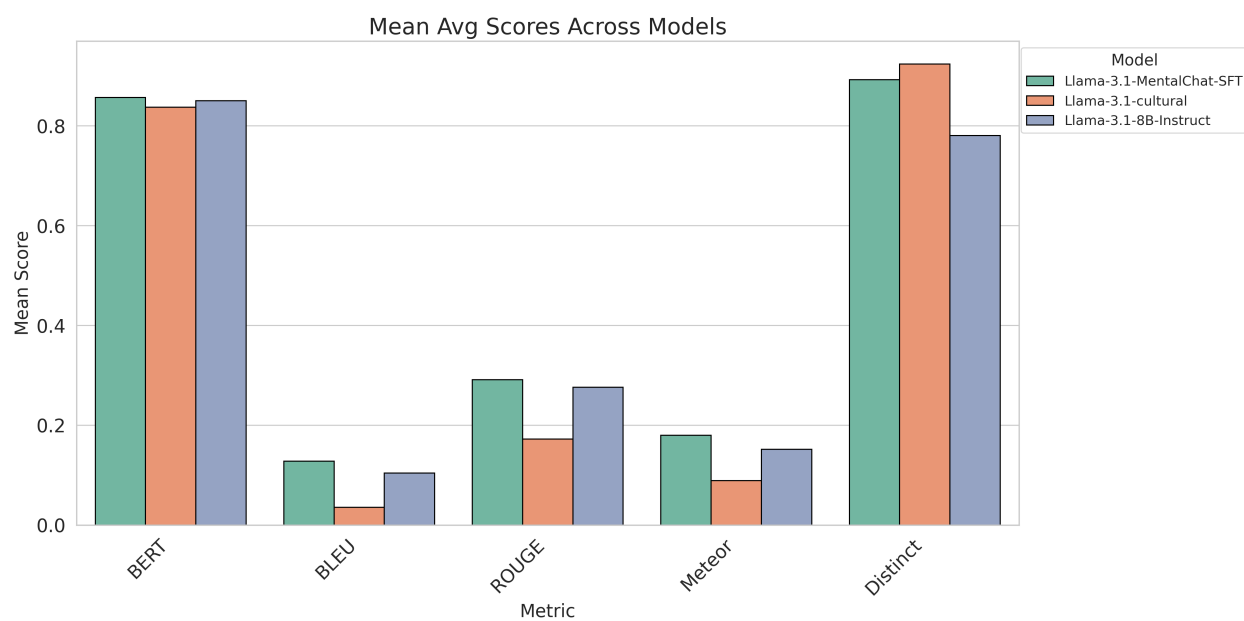


Figure 4.1: Mean Average scores across all the three Llama 3.1 models

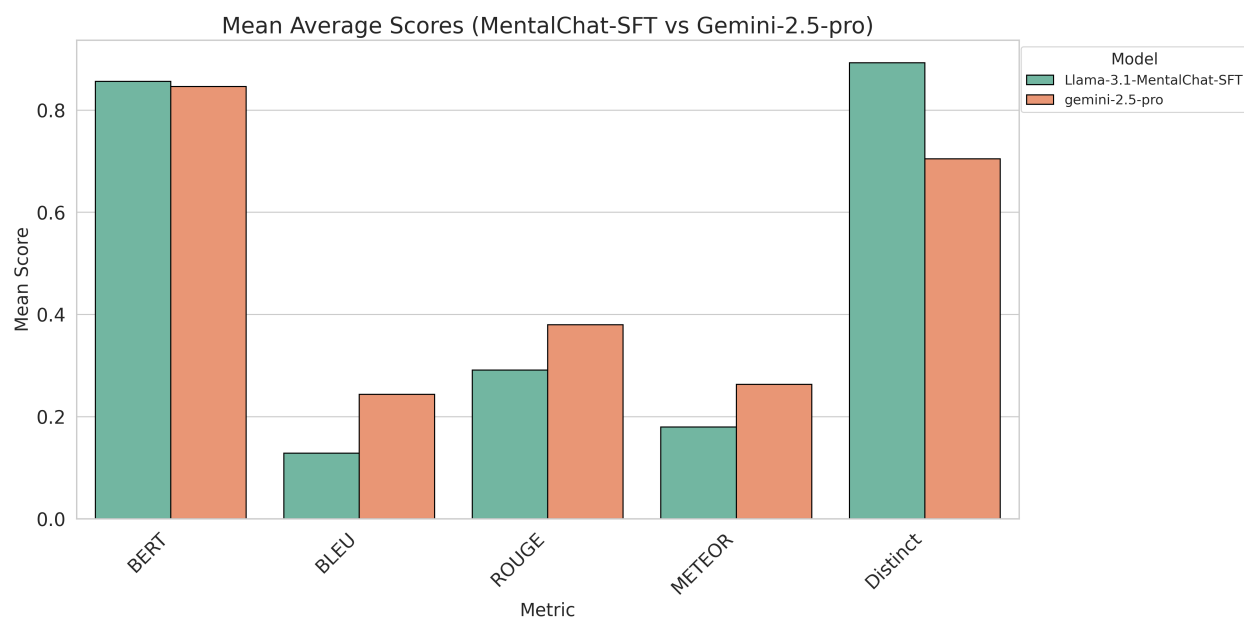


Figure 4.2: Mean Average scores comparing MentalChat-SFT and Gemini-2.5-Pro

the impact of my SFT strategy immediately apparent. The **Llama-3.1-MentalChat-SFT** model (green bars) consistently outperforms both the base instruct model (blue bars) and the culturally tuned variant (orange bars) across nearly all metrics relevant to quality and relevance (BERT, BLEU, ROUGE, and METEOR). This provides a clear, high-level confirmation that the fine-tuning process resulted in a model that is superior to its baselines in both semantic and lexical alignment with the target domain.

Figure 4.2 provides a comparison between our best-performing SFT model and the Gemini 2.5 Pro model. This plot effectively summarizes the core trade-off identified throughout our analysis. While the Gemini model can achieve superior performance on surface-level lexical matching, our SFT model demonstrates a more sophisticated and more valuable set of capabilities. Therefore, based on this comprehensive analysis, I select the **Llama-3.1-MentalChat-SFT** model to serve as the reasoning module for our LALM architecture. The fine-tuning process was demonstrably successful in adapting the model to the specific linguistic and semantic requirements of therapeutic dialogue.

4.2 Validation of Heuristic Thresholds

The analysis revealed a strong positive correlation, which is illustrated in the scatter plots presented in Figures 4.3 & 4.4.

Figure 4.3 displays human ratings of vocal energy on the x -axis (1–5 scale) against Gemini’s normalized energy scores on the y -axis. Each marker represents one utterance; the solid line is the least-squares fit with its 95% confidence band. The Pearson coefficient is $r = 0.53$, which falls in the conventional *moderate-positive* range ($0.50 \leq r < 0.70$). As the rating increases, Gemini’s estimate rises in parallel, and the spread of points narrows at the upper end, indicating greater model certainty for clearly high-energy speech.

Figure 4.4 shows the same analysis for pitch variability. Here $r = 0.52$, again a moderate, positive, and statistically reliable linear relation. The trend line confirms that utterances judged by humans as more expressive in pitch receive higher variability scores from Gemini, supporting the metric’s face validity.

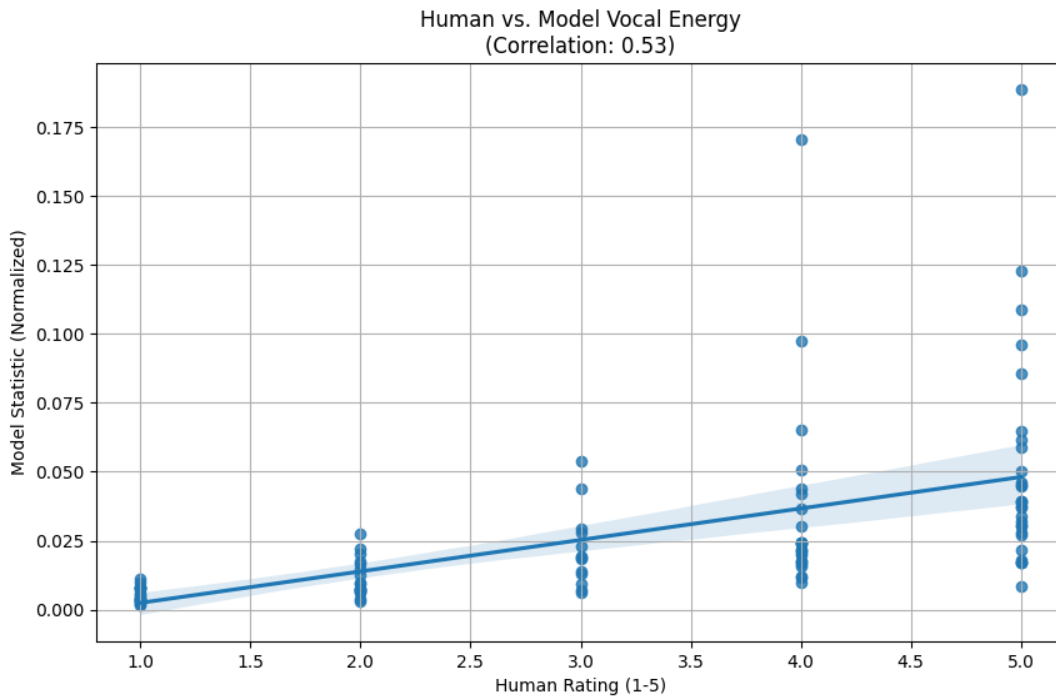


Figure 4.3: Correlation of scores for the vocal energy heuristic indicating positive trend.

Taken together, the two plots demonstrate that Gemini’s continuous acoustic scores track human perception with consistent, moderate strength across distinct prosodic dimensions. This empirical alignment justifies treating Gemini’s outputs as a sound proxy for expert annotation and using them to set data-driven thresholds. This validation allowed me to proceed with the Gemini-generated scores as a robust foundation for deriving my final thresholds. To derive these thresholds, I employed a quantile-based binning strategy on the continuous Gemini scores. For each heuristic, I calculated the 20th, 40th, 60th, and 80th percentiles to serve as the data-driven boundaries. The final quantile values used as thresholds were [2.0, 3.0, 3.0, 4.0] for vocal energy, [1.0, 2.0, 2.0, 4.0] for vocal expressiveness, and [2.0, 2.0, 2.0, 3.0] for speech rhythm.

It is important to acknowledge a limitation of this approach. While empirically derived,

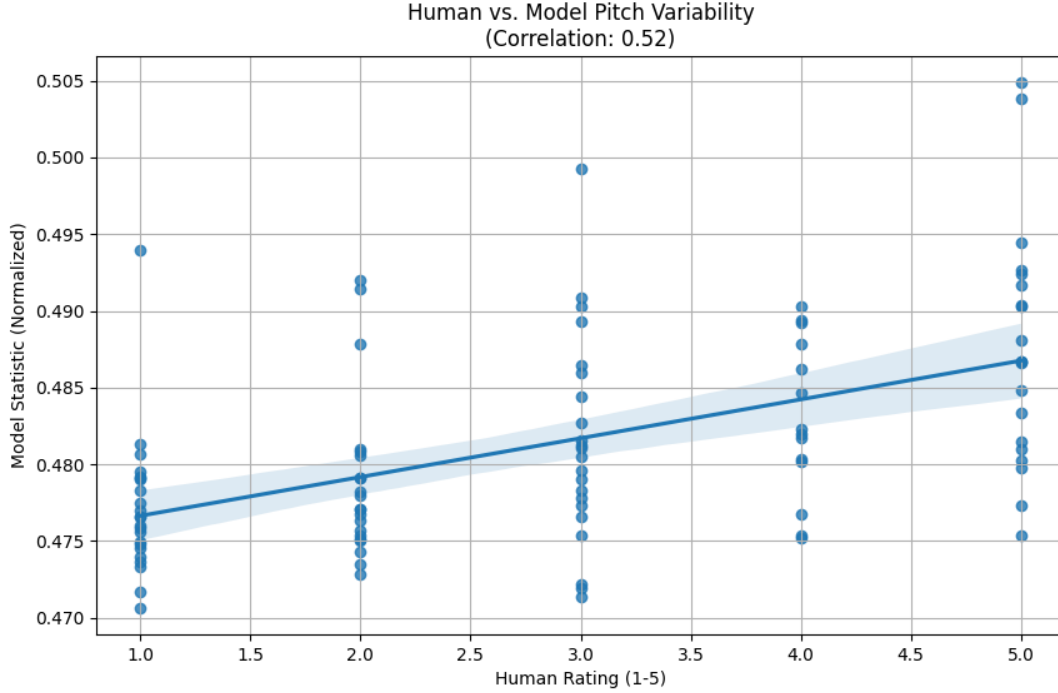


Figure 4.4: Correlation of scores for the pitch variability heuristic indicating positive trend.

these thresholds are ultimately grounded in the perceptual ratings of a single human annotator. Consequently, while the methodology is sound for this study, the accuracy and generalizability of these specific boundaries could be further enhanced by using a panel of multiple annotators. This process resulted in a set of empirically grounded, non-arbitrary thresholds for our heuristic model. Additionally, it would be insightful to gauge the practical effect of the proposed heuristic against a strong external baseline such as LTU-AS. Unfortunately, the public inference API for LTU-AS has been discontinued, and available computational resources were insufficient to replicate the model locally; a direct comparison therefore lies outside the scope of the present study.

4.3 Comparative Benchmarking and Model Selection

4.3.1 Benchmarking on Prepared DAIC-WOZ Dataset

As discussed in Section 3.5, I have conducted an evaluation of all five models on 345 prepared audio samples from the DAIC-WOZ dataset. The comparative evaluation of all five models yielded a clear performance hierarchy, as illustrated in Figure 4.5. This figure presents the aggregate normalized scores (0–100 scale) for each model across three separate evaluation runs, providing a comprehensive overview of their capabilities in audio understanding, generating empathetic and relevant responses from audio inputs.

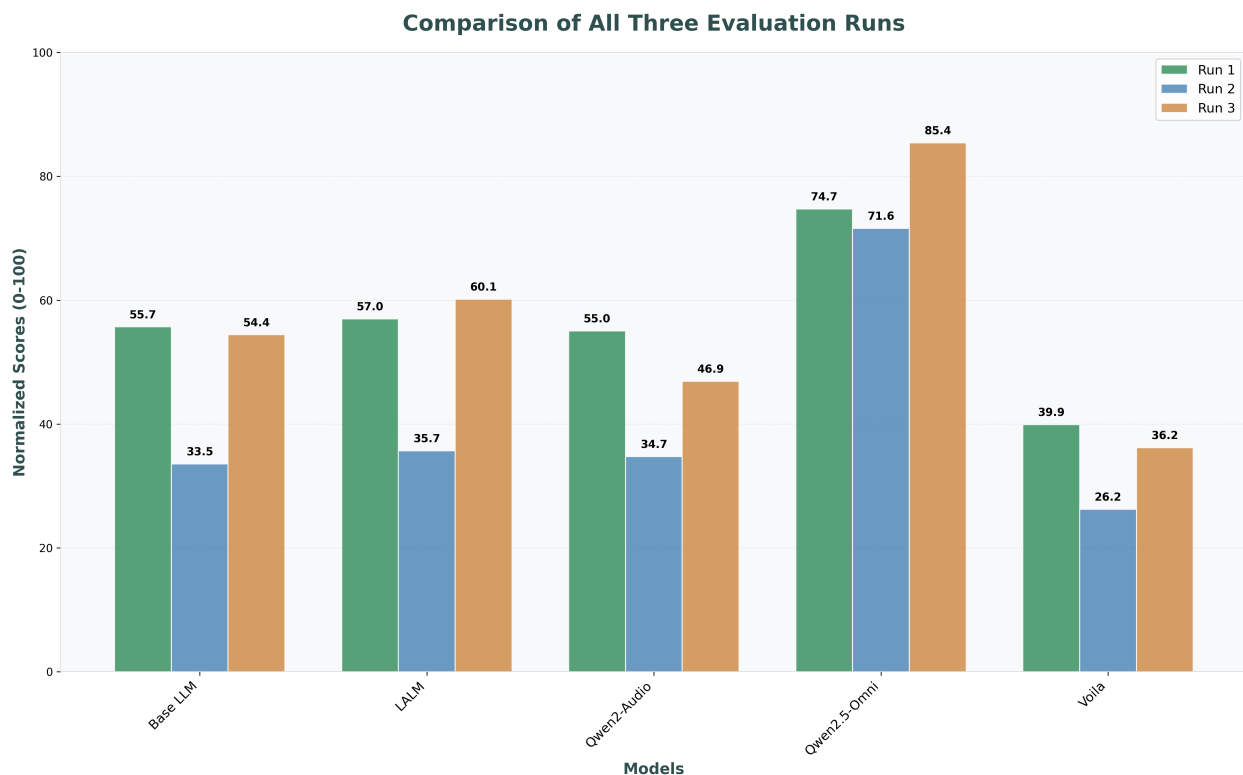


Figure 4.5: Comparison of all the three runs on DAIC-WOZ dataset

The most salient finding from this analysis is the clear dominance of the Qwen2.5-Omni model. Across all three evaluation runs, it achieved significantly higher scores than any other

model, with normalized scores of 74.7, 71.6, and a peak of 85.4. Even its lowest score in Run 2 surpasses the best performance of any other model in any run. This consistently superior performance underscores the impact of its advanced any-to-any architecture, which appears exceptionally well-suited for integrating the semantic and emotional nuances of spoken language to produce high-quality, relevant textual outputs based on its learned representations.

A direct comparison between the cascaded system and heuristic LALM reveals the incremental benefits of incorporating acoustic-level information. The heuristic LALM consistently outperformed the MentalChat-SFT model (ASR + SFT) in all three runs. This suggests that even a heuristic-based awareness of vocal cues and paralinguistic information provides valuable context that a purely text-based model, operating on ASR output alone, cannot access. This finding validates my hypothesis that paralinguistic cues are critical for empathetic response generation. These results show that my heuristic-based LALM approach is better at providing the text-LLM with audio information.

In contrast, the performance of the other end-to-end models, Qwen2-Audio-7B-Instruct and Voila-Chat, was notably weaker. The Qwen2-Audio model, despite its native ability to process audio and text, yielded scores (55.0, 34.7, 46.9) that were, at best, comparable to the cascaded system and did not demonstrate a clear advantage from its end-to-end design in this context. More strikingly, the Voila-Chat model, developed for voice-to-voice interaction, consistently registered the lowest performance, with scores of 39.9, 26.2, and 36.2. This was a truly unexpected result, and it suggests that while Voila may be optimized for low-latency, real-time interaction, its capacity for generating high-quality, therapeutically relevant text responses from voice is less developed than the other architectures evaluated.

A particularly insightful comparison can be drawn between the heuristic LALM and the Qwen2-Audio model. The LALM consistently outperformed Qwen2-Audio across all evaluation runs. In Run 1 and Run 2, the performance margin was modest, with LALM scoring 57.0 and 35.7, respectively, against Qwen2-Audio’s 55.0 and 34.7. However, the disparity became much more pronounced in Run 3, where the LALM achieved a score of 60.1, substantially higher than Qwen2-Audio’s 46.9.

This consistent performance advantage suggests that heuristic-based approach of the LALM is better suited for this specific task. Conversely, the Qwen2-Audio model, while being a native audio language model, processes the audio signal in a more integrated but implicit manner. Its internal fusion mechanisms must learn to disentangle and prioritize relevant acoustic features from the raw waveform without the aid of engineered heuristics. The results suggest that for the nuanced task of empathetic response generation, this implicit fusion is less effective than the LALM’s explicit feature extraction pipeline. The handcrafted acoustic descriptors appear to provide a more robust and salient signal for the LLM, leading to higher-quality responses.

Overall Score Frequency Analysis

Further nuance is added to the performance comparison by examining the score frequency distributions presented in Table 4.5. This table details the number of times each model received a specific rating from 1 (lowest) to 10 (highest) from the Gemini 2.5 Pro judge, providing a granular view of model consistency and response quality variance.

Model	Score Frequencies by Gemini 2.5 Pro									
	1	2	3	4	5	6	7	8	9	10
ASR + Llama-3.1-MentalChat-SFT	47	56	69	99	99	120	133	42	23	5
Heuristic LALM	29	54	60	99	78	95	131	65	21	9
Qwen2-Audio-7B-Instruct	84	73	64	93	83	50	66	59	101	6
Qwen2.5-Omni-7B	8	29	25	27	41	50	51	77	306	113
Voila-chat	88	173	93	81	63	53	45	32	16	4

Table 4.5: Frequency of Ratings (1-10) for Each Response Type Across All Runs

The data reveals a good contrast in the performance profiles of the models. The Qwen2.5-

Omni-7B model exhibits a distribution heavily skewed towards excellence. It received a remarkable 306 ratings of ‘9’ and 113 ratings of ‘10’, collectively accounting for nearly half of its total evaluations. These high scores indicate not only superior average performance but also responses of exceptional quality. For any real-world voice understanding tasks, particularly in sensitive domains like mental health, this level of consistency is paramount.

In the middle tier, the **ASR + Llama-3.1-MentalChat-SFT** and **Heuristic LALM** models display distributions centered in the 5-8 score range. Both models peak at a score of ‘7’ (133 and 131 frequencies, respectively) and show substantial frequencies across scores 4 through 8. This pattern suggests that these models reliably produce responses that are adequate or satisfactory but rarely achieve the level of excellence demonstrated by **Qwen2.5-Omni-7B**. They are less prone to catastrophic failures than the lowest-performing models but also lack the sophistication to consistently deliver top-tier, highly nuanced responses. The slightly higher frequency of ‘8’ scores for the **Heuristic LALM** (65 vs. 42) and ‘10’ scores (9 vs 5) correlates to the finding from the aggregate analysis that the inclusion of acoustic features provides a meaningful performance enhancement.

At the lower end of the performance spectrum, the **Voila-chat** model demonstrates a distribution heavily skewed towards poor quality. It received the highest frequency of ratings in the ‘1’ and ‘2’ categories (88 and 173, respectively), indicating that it frequently produced responses that were judged to be unhelpful and irrelevant. This failure is a shockingly unexpected result given **Voila-chat**’s voice-to-voice architecture, which one would presume to excel at such tasks. Similarly, the **Qwen2-Audio-7B-Instruct** model shows a concerning number of low scores, with its highest frequency of ‘9’ ratings (101) being undermined by a substantial number of ratings below ‘5’. This high variance in performance is an unexpected result for a native audio understanding model, suggesting a lack of reliability in consistently processing nuanced vocal cues.

In summary, the frequency analysis powerfully complements the aggregate scores by illustrating the performance consistency of each architecture. It confirms that the **Qwen2.5-Omni-7B** model’s superiority is not just due to a higher average but stems from its ability to reliably

generate high-quality outputs. Therefore, this analysis provided us a clear choice of the best performing model.

Analysis using ESHRO Framework

A detailed breakdown of model performance across the five ESHRO evaluation criteria—empathy, safety, helpfulness, relevance, and overall quality—is presented in Figure 4.6. This multi-dimensional analysis provides critical insights into the specific strengths and weaknesses of each architecture, moving beyond a single aggregate score to reveal a more nuanced performance landscape.

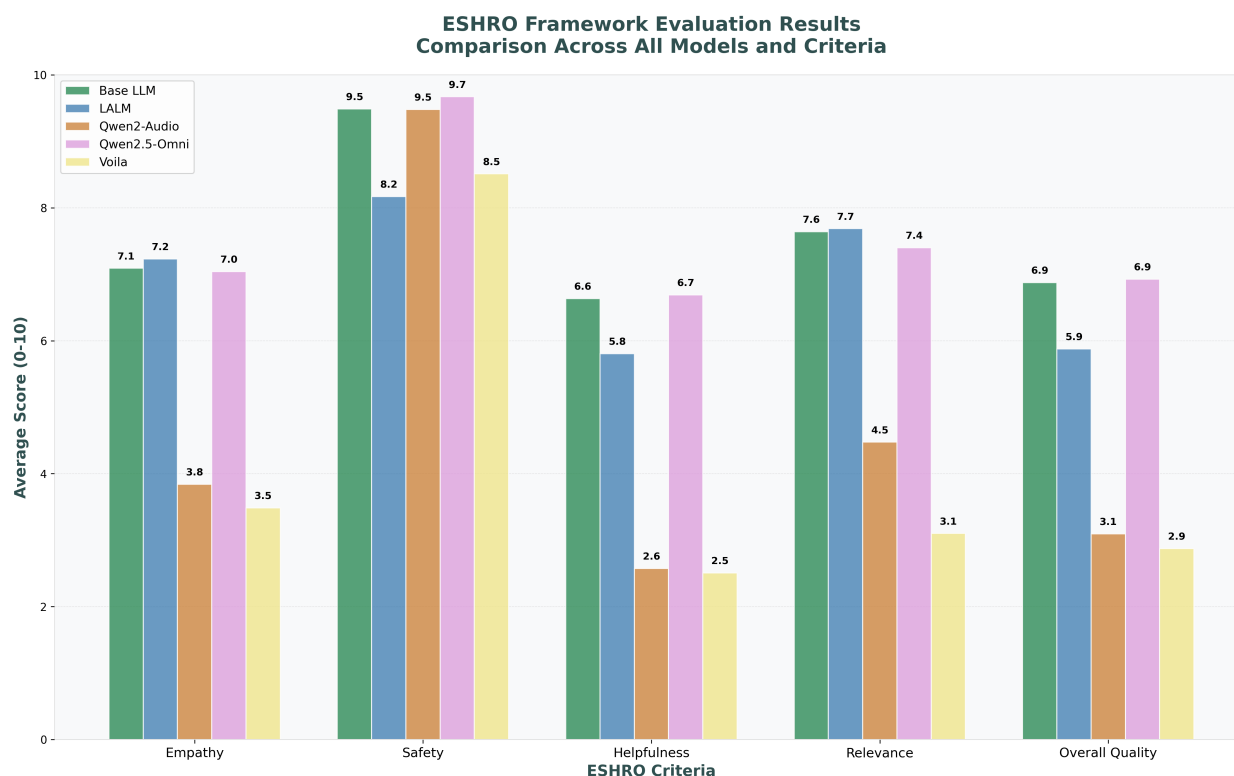


Figure 4.6: ESHRO Comparison of all the three runs on DAIC-WOZ dataset

On the core therapeutic dimensions of empathy and relevance, the models that were fine-tuned on domain-specific data—the Base LLM (7.1 empathy, 7.6 relevance) and the

Heuristic LALM (7.2 empathy, 7.7 relevance)—demonstrate a clear and decisive advantage. Their ability to generate responses that are perceived as both emotionally attuned and contextually appropriate is a direct result of their specialized training. The **Qwen2.5-Omni** model also performs strongly on these metrics (7.0 empathy, 7.4 relevance), indicating that its vast generalist knowledge can approximate domain expertise to a significant degree. In contrast, the native audio models, **Qwen2-Audio** and **Voila**, score very poorly (3.8 and 3.5 on empathy; 4.5 and 3.1 on relevance, respectively). This suggests that their architectures, while capable of processing audio, struggle to translate paralinguistic cues into nuanced, empathetic, and on-topic textual responses.

The safety metric reveals a different pattern, with all models achieving high scores. The **Qwen2.5-Omni** (9.7) and **Base LLM** (9.5) are nearly flawless, reflecting robust, well-established safety guardrails common in state-of-the-art foundation models. The **Voila** model’s respectable score of 8.5 in this category, despite its poor performance elsewhere, indicates that its safety alignment is a distinct and reasonably successful component of its design. The universally high scores suggest that ensuring response safety is a largely solved problem across these diverse architectures.

Helpfulness, which assesses the practical value of a model’s suggestions, proves to be a key differentiator. Here, the **Qwen2.5-Omni** (6.7) and the **Base LLM** (6.6) lead, indicating their ability to provide actionable guidance. The **Heuristic LALM** follows with a respectable score of 5.8. Again, the native audio models, **Qwen2-Audio** (2.6) and **Voila** (2.5), fail demonstrably, providing little to no practical help. The Overall Quality scores mirror this trend, with **Qwen2.5-Omni** and the **Base LLM** tied at 6.9, followed by the **Heuristic LALM** at 5.9. The catastrophic failure of **Qwen2-Audio** (3.1) and **Voila** (2.9) on this composite metric highlights their current unsuitability for this application.

Conclusion

The comprehensive evaluation, combining the ESHRO framework analysis with the score frequency distributions, leads to a decisive conclusion. While domain-specific fine-tuning,

as seen in the **Base LLM** and **Heuristic LLM**, is critical for achieving high empathy and relevance, it is not sufficient on its own. The **Heuristic LLM**, despite its innovative use of acoustic features, is hampered by being the least safe model and by its mediocrity in providing helpful responses.

In contrast, the **Qwen2.5-Omni** model demonstrates outstanding and, crucially, consistent performance across all key dimensions. It is the safest model, one of the most helpful and relevant, and achieves a top-tier overall quality score. This exceptional balance, combined with the score frequency analysis, showed its responses are consistently rated in the highest ‘9’ and ‘10’ categories. Its ability to excel without domain-specific fine-tuning points to a more advanced and powerful underlying architecture for voice understanding and reasoning. Therefore, the **Qwen2.5-Omni** model is selected as the best-performing model. It is the most suitable and promising foundation for developing emotionally intelligent voice language models.

4.3.2 Benchmarking on Voila-Bench

As per the discussion in Section 3.5.4, I present and analyze the results of Voila-Bench evaluation. Following the initial evaluation on the DAIC-WOZ dataset, I conducted a more standardized assessment using the VoilaBench benchmark to rigorously test the models’ voice understanding and reasoning capabilities across a diverse set of general knowledge domains. The results from this benchmark provide a clear and definitive picture of the models’ performance hierarchies.

The official Voila-Bench dataset consists of 1,400 audio-text question-answer pairs. For our benchmarking, I selected Gemini 2.5 Pro as the judge over GPT-4o, citing its state-of-the-art performance and high LLM leaderboard rankings (as of June 2023). All model responses were generated on a single NVIDIA RTX A6000 GPU with 48GB of VRAM. Due to API rate limits for Gemini 2.5 Pro, my analysis was ultimately limited to a subset of 100 samples from the full Voila-Bench dataset.

Overall Model Performance on VoilaBench

The aggregate performance of the three models, as evaluated by Gemini 2.5 Pro and presented in Figure 4.7, reveals a performance differential. The **Qwen2.5-Omni** model emerges as the highest scorer, achieving an average score of 61.8 out of 100. This result is more than 1.6 times higher than its closest competitor, **Qwen2-Audio**, which scored 37.4, and a staggering 3.5 times higher than **Voila**, which registered the lowest score of 17.5.

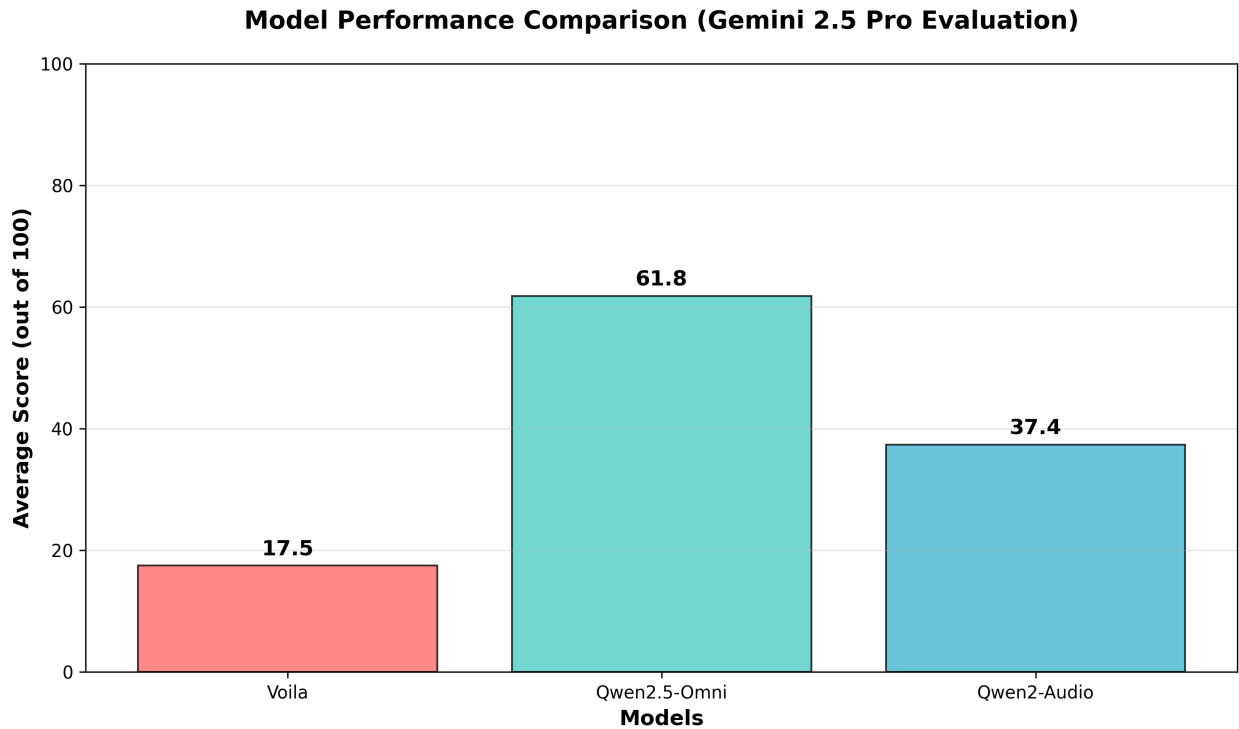


Figure 4.7: Overall model performance comparison on the VoilaBench benchmark

This wide performance gap reinforces the findings from our initial DAIC-WOZ and ESHRO evaluations. The superior result of **Qwen2.5-Omni** on a standardized, multi-domain benchmark confirms that its architectural strength is not limited to a specific task but represents a more general and robust capability for voice-based reasoning. The underwhelming performance of **Qwen2-Audio** and the near-complete failure of **Voila** are particularly re-

vealing. Despite being advanced, native audio-understanding architectures, they lack the sophisticated reasoning or knowledge base required to compete with a top-tier any-to-any model on complex question-answering tasks. This suggests that the ability to process raw audio is only a preliminary step; the core challenge lies in the subsequent semantic interpretation and reasoning, where **Qwen2.5-Omni** excels.

Analysis of Score Distributions

A deeper analysis of the score distributions, shown in Figure 4.8, further illuminates the nature of each model’s performance. The plots reveal not just average performance but also the consistency and reliability of the models’ outputs.

- **Qwen2.5-Omni:** The score distribution is heavily skewed to the right, indicating a high frequency of excellent responses. The most prominent peak is at a score of ‘10’, with 35 instances rated as “perfect.” A substantial portion of its responses fall within the 7-10 range. This distribution signifies a model that is not only highly capable but also reliable, consistently producing high-quality answers. Its mean score of 6.18 accurately reflects this strong, positive performance profile.
- **Qwen2-Audio:** This model’s distribution reveals a high degree of performance variance and unreliability. It exhibits a bimodal-like pattern, with its largest peak at the lowest possible score of ‘1’ (47 instances), but also a smaller, secondary peak at a score of ‘9’. This indicates that the model is highly inconsistent, capable of generating both extremely poor and very good responses. Such unpredictability makes it ill-suited for any application requiring dependable performance.
- **Voila:** The distribution for the Voila model confirms its failure on this benchmark. The vast majority of its scores are clustered at the lowest end of the scale, with a massive peak at a score of ‘1’ (75 instances). With a mean score of only 1.75, the model demonstrates a near-total inability to correctly answer the questions posed in

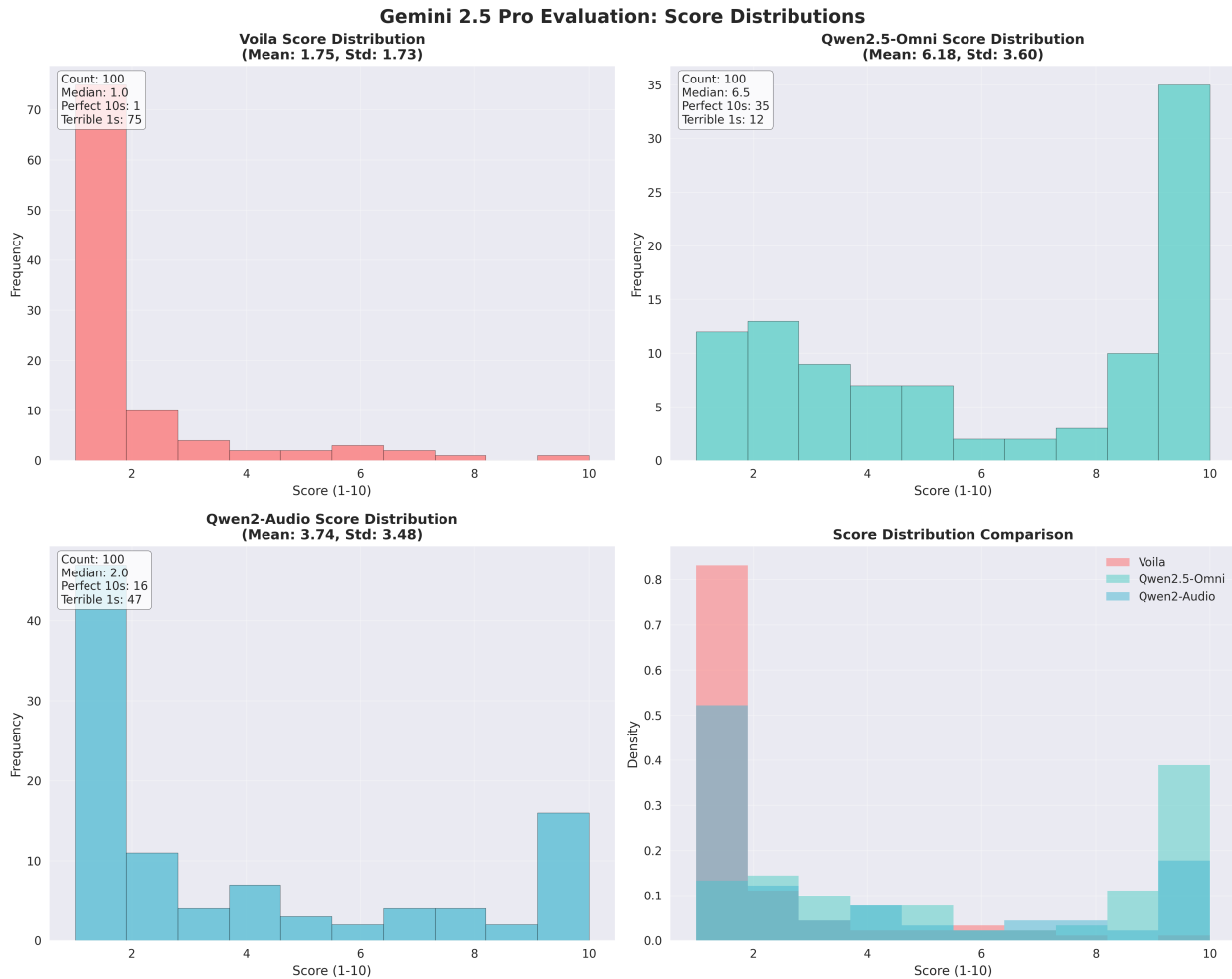


Figure 4.8: Score distributions for Voila, Qwen2.5-Omni, and Qwen2-Audio on the VoilaBench benchmark

the benchmark. This result is especially surprising and disappointing, given that this evaluation was conducted on its own namesake benchmark, which was presumably designed to highlight its strengths.

The stacked density plot in the bottom-right of Figure 4.8 provides a visual summary of these findings. It clearly shows the Voila and Qwen2-Audio models dominating the low-score region, while the Qwen2.5-Omni model almost exclusively occupies the high-score territory.

This confirms that the performance differences are not marginal but represent fundamental gaps in capability and reliability.

Considering all the evidence from the DAIC-WOZ, ESHRO framework evaluation and the VoilaBench benchmark, its a clear conclusion that the Qwen2.5-Omni model has demonstrated itself to be the best-performing model in terms of both voice understanding and the ability to reason over human voice. Its consistent, top-tier performance across diverse datasets and evaluation criteria, coupled with its high reliability as shown in the score frequency analyses, makes it the highest-rated model for selection in any application requiring advanced and dependable voice-based reasoning.

4.3.3 Analysis of VocalBench Results

To see how the models compare to the best in the field, I analyzed the results from the VocalBench paper (Liu et al., 2025). This paper presents a tough and fair test for voice AI models, measuring a wide range of skills including language understanding, voice clarity, and conversational ability. It uses many different scores to check how well a model performs in areas like reasoning, emotion, and safety. This allows for a direct comparison of the top models. My analysis of this data confirms the specific strengths of the Qwen2.5-Omni model.

The VocalBench results in Table 4.6 clearly show that while some models are good at specific things, Qwen2.5-Omni is a powerful and well-rounded performer, proving it is the best in the most important areas of voice understanding and reasoning. A model’s ability to understand and think is its core intelligence. In this area, Qwen2.5-Omni is the clear winner, achieving the best reasoning score of all 12 models tested. This smart reasoning is supported by its very high knowledge accuracy, meaning it has a large and correct base of information to draw from. This combination of knowledge and reasoning is essential for any advanced AI and makes Qwen2.5-Omni stand out.

The tests on voice quality also confirm the model’s excellent ability to process human speech. It has the best clarity score, making the fewest mistakes when turning speech into text. This is a basic but critical skill for any voice AI. It also scores high on Fluency, meaning

Model	Size	Semantic			Acoustic		Chat						Robust	Overall
		Knowledge	Reasoning	Creativity	Fluency	Clarity	Single Round	Multi Round	IF	Emotion	Safety	Latency		
		Accuracy (%)	Score	Score	UTMOS	WER (%)	Score	Score	FR (%)	Score	RR (%)	RTF		
Mini-Omni	0.5B	2.20	1.291	1.4725	4.435	19.571	1.645	-	0.00	5.428	81.25	0.3781	84.14	40.646
Mini-Omni2	0.5B	4.65	1.501	1.8025	4.413	36.269	1.915	-	0.11	5.709	88.50	0.2001	82.26	43.224
SLAM-Omni	0.5B	12.05	1.875	2.5175	4.424	6.065	2.880	1.9800	3.11	6.452	90.25	0.4925	77.91	54.649
VocalNet-1B	1B	43.00	2.869	3.1800	4.437	5.123	3.335	3.2550	16.11	6.754	89.00	0.1632	92.42	66.632
LLaMA-Omni	8B	37.40	2.591	2.8475	3.959	2.842	3.300	3.1525	14.89	6.128	27.75	0.0958	83.59	57.107
Freeze-Omni	7B	44.25	3.530	2.8850	4.381	11.460	2.960	-	12.05	6.164	86.50	0.2618	65.25	58.362
Baichuan-Omni-1.5	7B	49.85	3.770	3.5900	4.014	23.452	3.840	-	28.89	5.424	83.00	1.4900	74.85	60.239
GLM-4-Voice	9B	56.40	3.641	3.2900	3.869	11.565	3.615	3.7300	31.67	6.904	71.50	0.7870	57.10	61.388
Kimi-Audio	7B	62.15	3.132	3.0950	2.360	38.001	3.150	3.5350	48.59	6.838	83.75	0.7331	93.20	62.382
MiniCPM-o 2.6	7B	70.00	3.648	3.3550	4.054	18.735	3.165	3.6675	30.00	7.080	83.25	0.4509	87.27	63.886
Qwen2.5-Omni	7B	69.50	4.361	3.1825	4.174	1.154	3.538	4.0125	27.00	6.386	71.75	1.7243	91.86	73.327
VocalNet-8B	8B	67.95	3.748	3.5050	4.449	4.686	3.530	3.9175	35.89	7.117	92.25	0.2496	92.66	74.639

Table 4.6: Evaluation performance of 12 models on VocalBench. Bold indicates the optimal result in each subgroup. Source: (Liu et al., 2025).

its responses sound natural. Furthermore, a voice model must be good at handling back-and-forth conversations. Here too, Qwen2.5-Omni shows its skill by getting the highest score for multi-round conversations. This shows it can keep track of long, complex discussions and have meaningful interactions, which is crucial for mental health therapy.

Most importantly, its conversation skill is combined with an understanding of emotion. The model earned a strong emotion score without needing any special training for it. This suggests it has a natural ability to understand and respond to feelings, simply as part of its powerful voice understanding and reasoning skills. This is different from other models that may need to be specifically trained to detect emotion.

In conclusion, the VocalBench results offer clear proof that supports my earlier findings. While other models may be good at one or two things, Qwen2.5-Omni is the only one that performs at a top level across all the most critical and difficult categories: reasoning, voice understanding, and multi-round conversation. Its excellent results in these key areas, along with its built-in emotional awareness, confirm its position as the best and most flexible voice-language model. It is a complete end-to-end model that understands, reasons, and talks with

a high level of skill and is emotionally intelligent by nature. This makes Qwen2.5-Omni the clear choice for my further fine-tuning, and any task that requires the highest level of voice intelligence.

4.3.4 User Survey Analysis

To complement our objective benchmark evaluations, I conducted a direct comparative user study, as discussed in Section 3.6, to assess the perceived quality of the two most architecturally distinct models: Model A (Qwen2.5-Omni) and Model B (Voila-Chat). This study provides critical, human-centered data on how users experience and rate the outputs of these models in the real world. The results, presented below, offer evidence of user preference and validate our earlier findings.

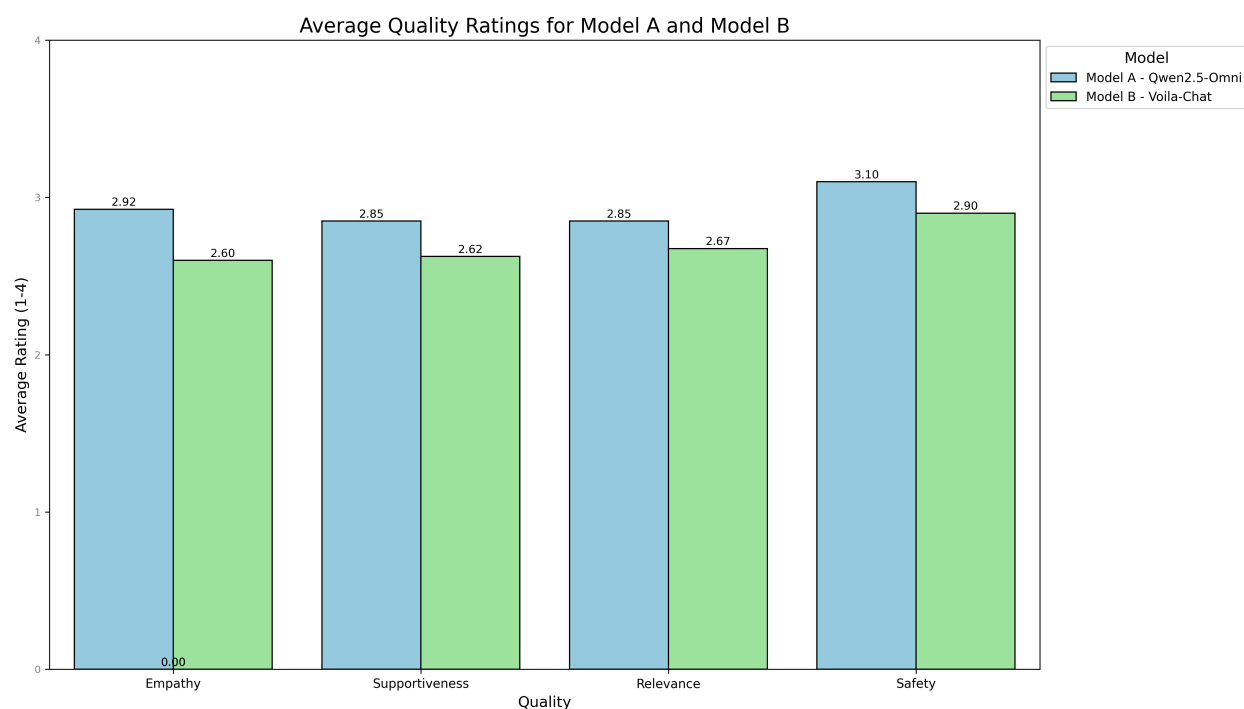


Figure 4.9: Average quality ratings on a 1-4 scale for Model A (Qwen2.5-Omni) and Model B (Voila-Chat) across four key criteria.

As of 08-04-2025, I have a total of 49 users who participated in this study. Participants were asked to rate each model's response on a 1-4 scale across four qualitative criteria adapted from the ESHRO framework: Empathy, Supportiveness, Relevance, and Safety. As illustrated in Figure 4.9, Model A (**Qwen2.5-Omni**) consistently and clearly outperformed Model B (**Voila-Chat**) on every single metric. Specifically, Model A achieved higher average scores in empathy (2.92 vs. 2.60), supportiveness (2.85 vs. 2.62), and relevance (2.85 vs. 2.67). This indicates that users perceived its responses as more emotionally attuned, more encouraging, and more contextually appropriate. The superior performance in these core therapeutic dimensions highlights Model A's advanced ability to not only understand the explicit content of the user's speech but also to grasp the underlying emotional and contextual nuances. Furthermore, Model A was also rated as safer (3.10 vs. 2.90), suggesting that its responses were perceived as more reliable and less likely to be dismissive or harmful.

The superiority of Model A becomes even more pronounced when examining the overall quality ratings (on a 1-10 scale) for each of the 10 distinct audio clips used in the study. Figure 4.10 shows that Model A received a higher average rating than Model B in 9 out of the 10 scenarios.

More than just the higher average, the consistency of Model A's performance is a key finding. Its scores are clustered tightly at the top of the scale, ranging from a low of 7.7 to a high of 8.8. In contrast, Model B's performance is far more erratic, with scores ranging from a low of 5.8 to a high of 8.2. This high variance suggests that while Model B is occasionally capable of producing a good response, it is fundamentally unreliable. Model A, on the other hand, is a consistently high performer, which is a critical attribute for any system intended for real-world deployment, especially in a sensitive application like mental health support.

As shown in Figure 4.10, if I look at overall user preference across all metrics, Model A achieved approximately 2:1 superior performance compared to Model B, with winning percentages ranging from 50.5% to 55.2%. In evaluating which model more accurately understood key points and details from audio content, Model A achieved a 50.5% win rate compared to Model B's 25.0%, representing a 25.5 percentage point advantage. This perfor-

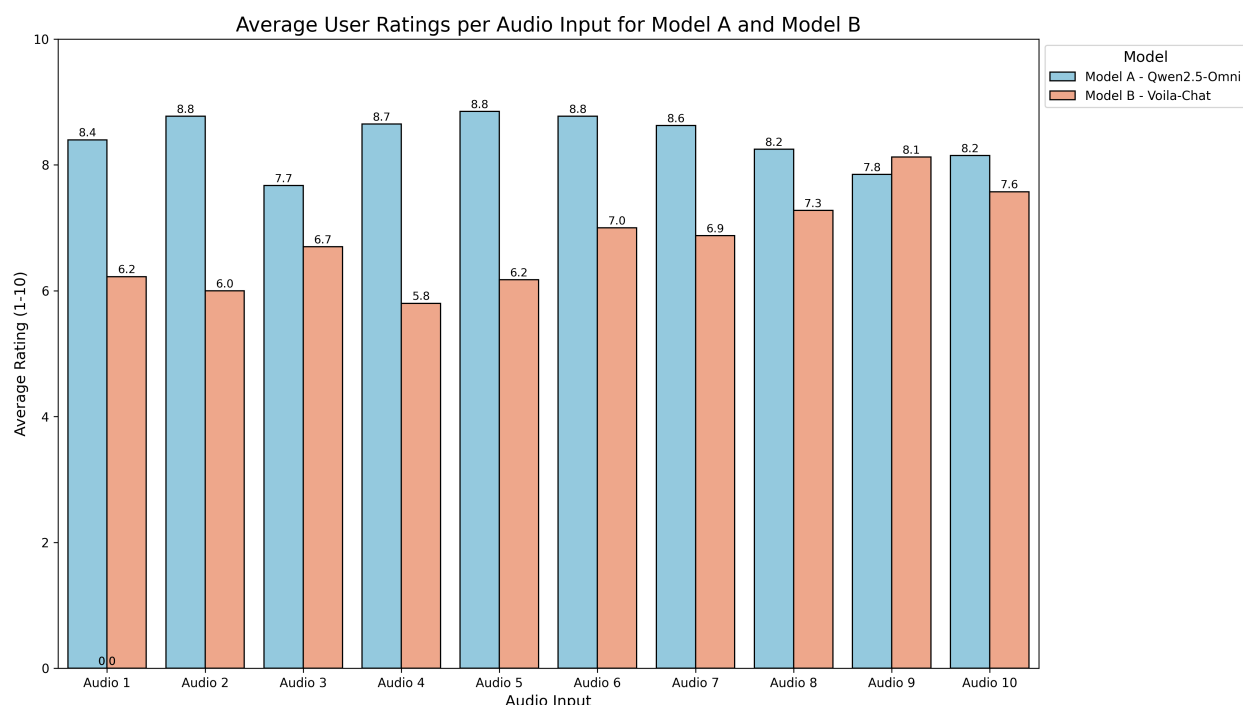


Figure 4.10: Average user ratings (1-10) per audio input for Model A (Qwen2.5-Omni) and Model B (Voila-Chat)

mance differential is statistically significant with a medium effect size (Cohen's $h = 0.534$).

The analysis across 10 individual questions revealed consistent Model A superiority, with notable exceptions in Questions 3 and 9, where performance was more competitive. In Question 9 specifically, Model B achieved a rare victory with 47.5% vs. Model A's 27.5%, suggesting certain audio scenarios may favor Model B's approach. Importantly, 23.2% of responses indicated both models performed equally well, suggesting that while Model A generally excels, there are scenarios where both models achieve comparable understanding of key content details.

In conclusion, across every qualitative metric and in nearly every conversational scenario, user evaluation demonstrated a strong and consistent preference for the responses generated by Model A (Qwen2.5-Omni). Users found its responses to be more empathetic, supportive,

relevant, and safe. They also rated its overall quality significantly higher and more consistently than its competitor. This human-centered evaluation powerfully corroborates the findings from my objective benchmarks, confirming that **Qwen2.5-Omni** is not just technically proficient but also excels in the nuanced, human aspects of conversation. Its ability to reliably generate high-quality, emotionally intelligent responses makes it undoubtedly the best-performing model for advanced voice understanding and reasoning.

4.4 Evaluation of DPO Model

4.4.1 *Metric-Based Evaluation*

This section presents a detailed comparative evaluation of the baseline Qwen2.5-Omni model and its DPO-aligned variant, focusing on quantitative metrics that reflect semantic fidelity, linguistic coherence, and overall textual quality.

BERT Scores: Precision, Recall, and F1

The BERT-based metrics, displayed in Figure 4.11, reveal the nuanced impact of DPO alignment on the model’s textual outputs. The DPO-aligned model registers consistent improvements across all three key BERT-derived scores: a precision increase from 0.86 to 0.88, a recall rise from 0.85 to 0.88, and an F1 enhancement from 0.85 to 0.88. These gains indicate that DPO fine-tuning leads to generated responses that are not only more accurate but also more comprehensive in capturing relevant content, resulting in better alignment with reference responses.

BLEU Scores: N-gram Precision

The BLEU metric evaluation, focusing on 1-gram and 2-gram overlaps and displayed in Figure 4.12, corroborates the BERT findings and further accentuates the benefits of DPO fine-tuning. The DPO model achieves a nearly 44% relative increase in the 1-gram BLEU score (from 0.23 to 0.33) and a 67% relative increase in the 2-gram BLEU score (from 0.12 to

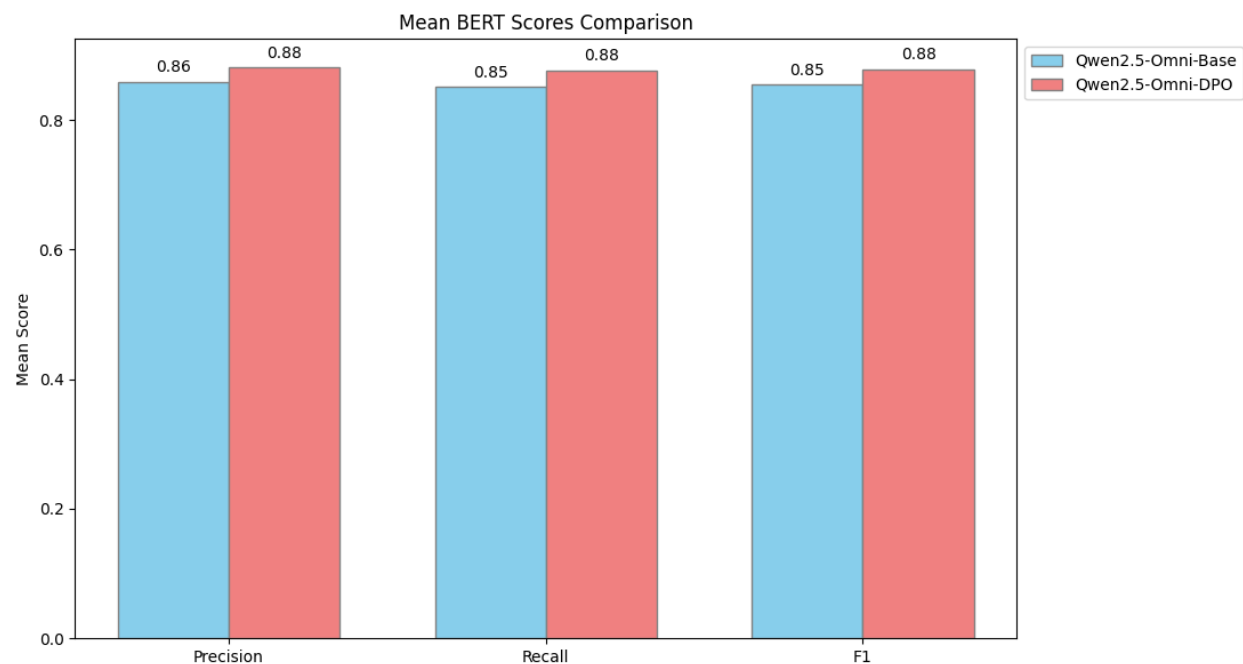


Figure 4.11: Mean BERT Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.

0.20). Such improvements underscore enhanced fluency and contextual coherence, suggesting that DPO alignment better equips the model to generate text sequences that closely mirror natural language patterns seen in ground truth data.

ROUGE Scores: Overlap-Based Recall Measures

ROUGE scores, displayed in Figure 4.13, also depict a significant positive shift post-DPO alignment. The DPO model surpasses the baseline in all ROUGE sub-metrics: ROUGE-1 rises from 0.339 to 0.447, ROUGE-2 from 0.092 to 0.16, and ROUGE-L from 0.178 to 0.229. These increments reflect improved recall of salient phrases and syntactic structures, thereby enhancing the informativeness and cohesiveness of generated replies.

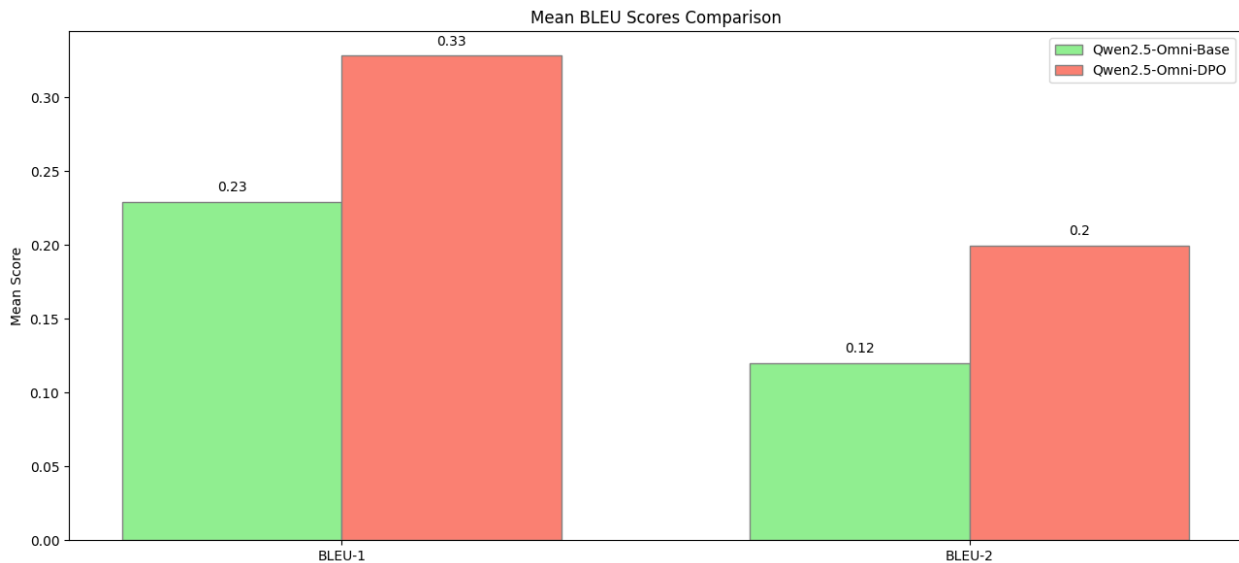


Figure 4.12: Mean BLEU Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.

METEOR Scores: Semantic and Structural Alignment

The METEOR score, which provides a more sophisticated measure of semantic similarity by incorporating synonymy and stemming, reveals a substantial performance gain after DPO alignment. As shown in Figure 4.14, the score increased by over 41%, from 0.213 for the baseline model to 0.301 for the DPO-aligned version. This significant improvement indicates that the DPO model is not merely matching surface-level n-grams but is generating responses that are semantically closer to the reference translations. It successfully captures the intended meaning, even when using different but equivalent wording, which is a hallmark of a more advanced and human-like language understanding.

Perplexity Scores: Language Model Confidence

Perplexity serves as an inverse measure of a language model's confidence; a lower score indicates that the model is less "surprised" by the sequence of words in the test set, which

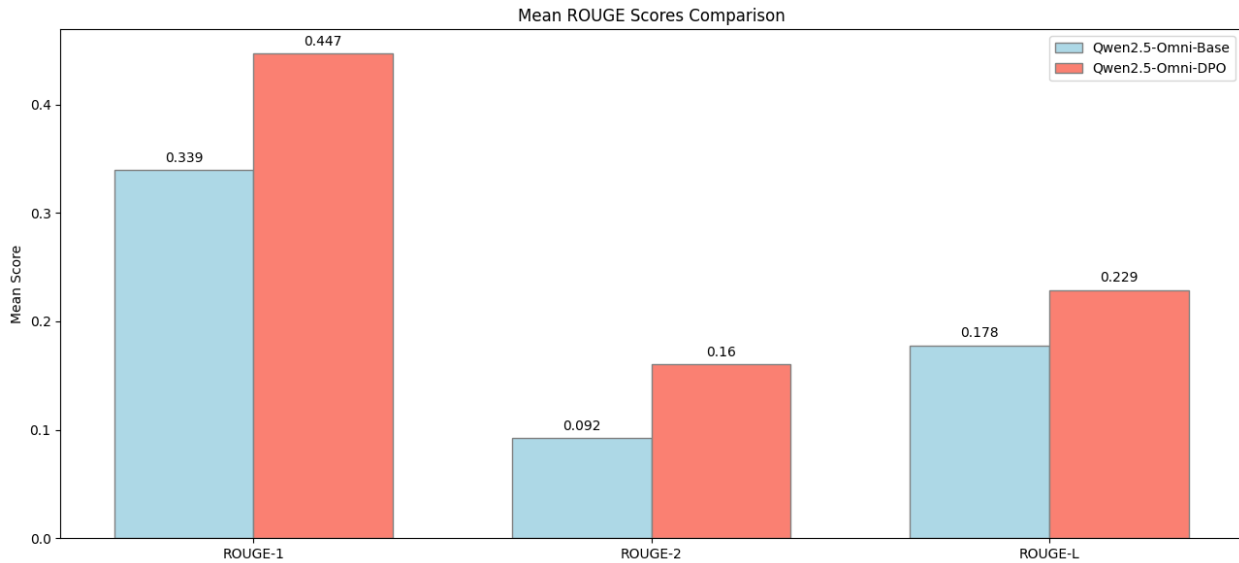


Figure 4.13: Mean ROUGE Scores Comparison between the base Qwen2.5-Omni model and its DPO-aligned version.

translates to higher fluency and better modeling of the underlying language distribution. The results in Figure 4.15 show a clear, albeit modest, reduction in perplexity from 11.7 for the base model to 10.7 for the DPO model. This decrease signifies that the DPO alignment process has refined the model’s internal probability distributions to better match those of high-quality, human-preferred text, resulting in more coherent and predictable language generation.

Distinct-n Scores: Lexical Diversity

The Distinct-n scores measure the lexical diversity of the generated text by calculating the ratio of unique n-grams to the total number of n-grams. As seen in Figure 4.16, the results are nuanced but telling. The Distinct-1 score, measuring the variety of individual words, remains largely unchanged (0.55 vs. 0.541), suggesting that DPO does not simply force the model to use a wider vocabulary. However, the Distinct-2 score, which measures the variety

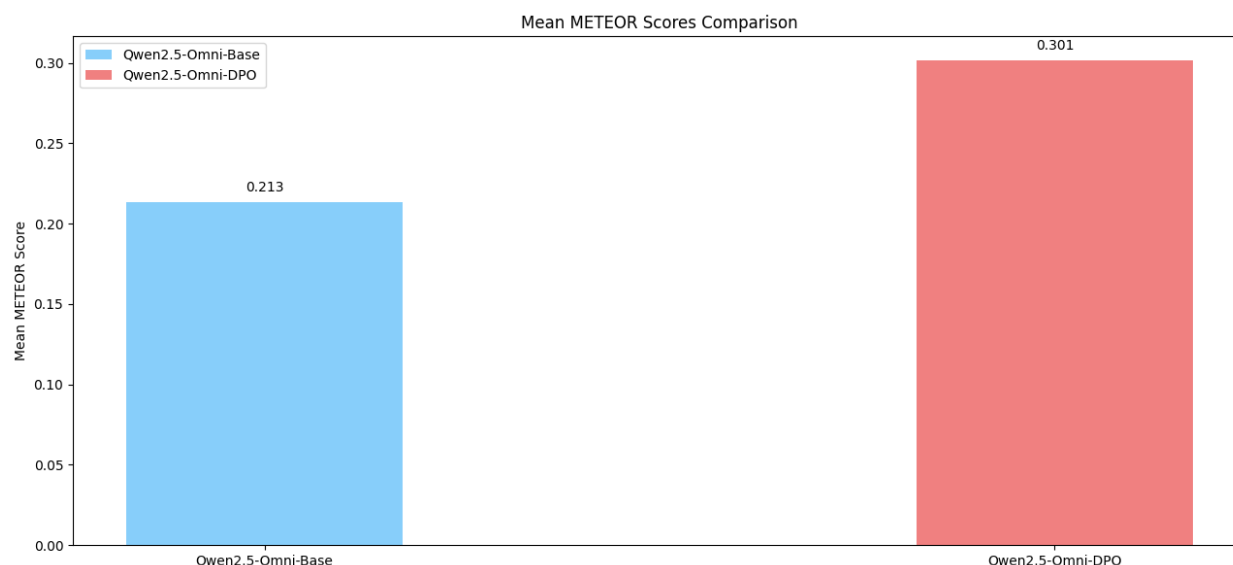


Figure 4.14: Mean METEOR Scores Comparison. The DPO-aligned model shows a significant improvement in semantic alignment.

of two-word phrases (bigrams), shows a significant increase from 0.819 to 0.877. This is a critically important finding. It indicates that the DPO-aligned model is more creative and less repetitive in its phrasing and sentence construction. It excels at generating more varied and engaging language, which is essential for avoiding a robotic or canned feel in conversational AI.

Together, these quantitative evaluations compellingly demonstrate that DPO alignment benefits the Qwen2.5-Omni model, producing responses that are semantically accurate (BERT, METEOR), linguistically fluent and coherent (Perplexity, BLEU), contextually rich (ROUGE), and, critically, more lexically diverse and engaging (Distinct-2). This signifies that the DPO fine-tuning of Qwen2.5-Omni on mental health text data fosters a deeper alignment with human preferences and natural language patterns. Such evidence strongly supports the incorporation of DPO-based training regimes in advancing the models in emotional intelligence and therapeutic response generation.

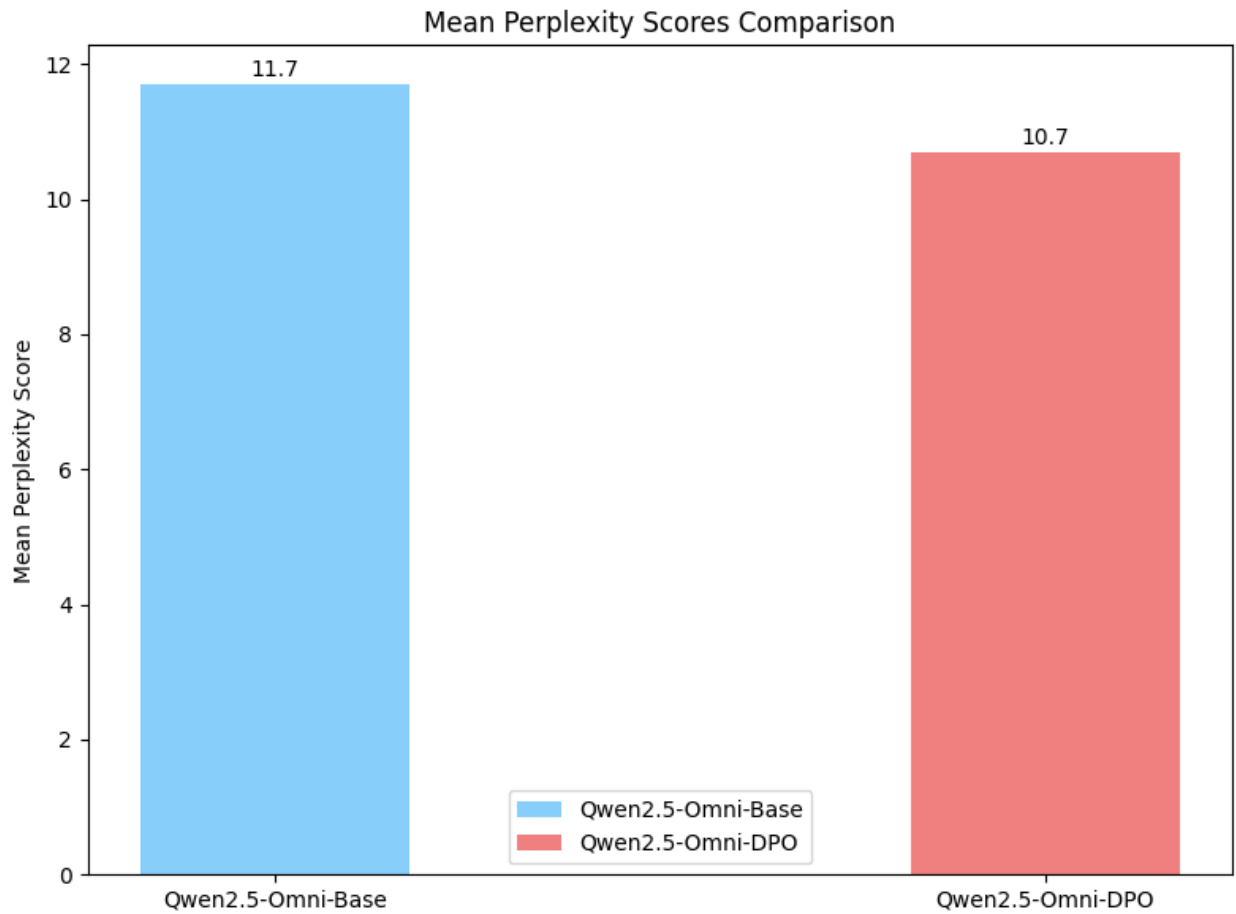


Figure 4.15: Mean Perplexity Scores Comparison. Lower perplexity for the DPO model indicates improved fluency and model confidence.

4.4.2 Human Evaluation

The results of the human expert evaluation revealed clear and consistent preferences for the DPO model over the base model across several key performance dimensions. Expert evaluators, who have professional backgrounds in therapy and mental health, provided ratings on various qualitative and quantitative aspects after listening to paired voice responses generated from 11 audio clips. To ensure fairness and minimize order effects, the audio samples were shuffled during the evaluation. As of 08-11-2025, the models have been evaluated by

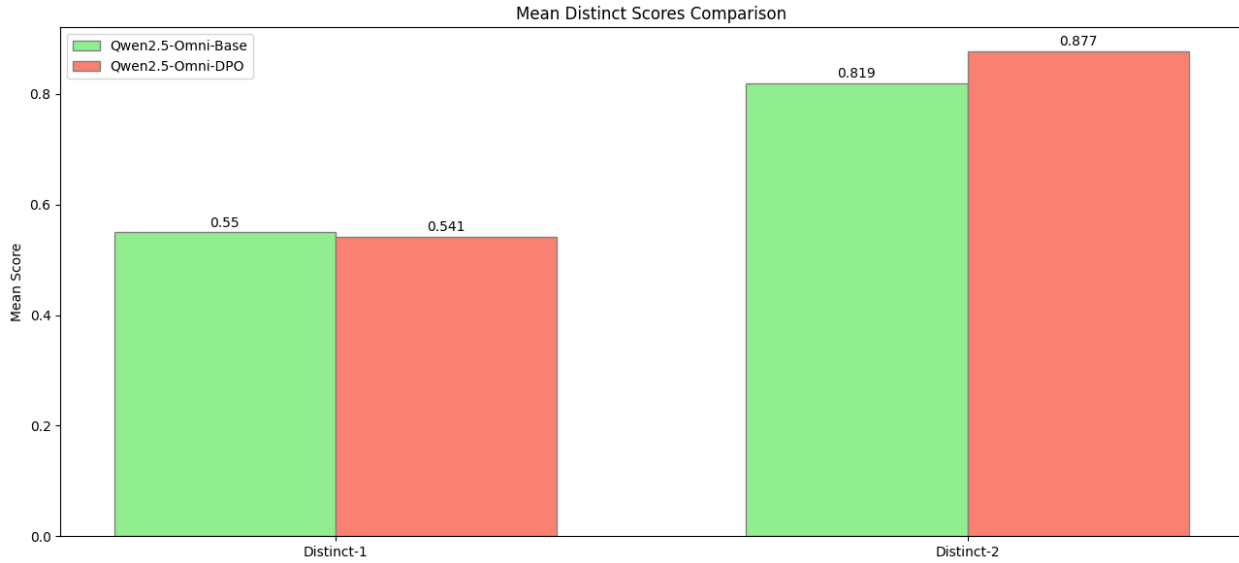


Figure 4.16: Mean Distinct Scores Comparison. The DPO model shows a notable increase in bigram diversity (Distinct-2), indicating more varied phrasing.

one human expert.

Overall, as shown in Table 4.7, the DPO model was rated higher in its ability to perceive and reflect emotions within the user’s speech. Experts noted that the DPO model more accurately captured the subtle emotional nuances conveyed in the audio, thereby producing responses that felt more genuinely empathetic and emotionally aligned. This enhanced emotional resonance was reflected not only in direct emotion perception scores but also in the model’s demonstrated compassion and professionalism, which closely matched expert expectations of therapeutic communication.

In terms of contextual understanding and personalization, the DPO model again outperformed the base model. Evaluators observed that DPO responses reflected a deeper comprehension of users’ needs and contexts, translating into more tailored and relevant conversational exchanges. These strengths in reasoning and personalization suggest the benefits of the DPO training process in fine-tuning model responsiveness.

Evaluation Question	Base Model	DPO Model
Overall Better Model		✓
Better Emotion Perception		✓
Better Context Understanding	✓	✓
Stronger Reasoning		✓
More Like Professional Therapist		✓
More Natural	✓	
Empathy	Good (4/5)	Good (4/5)
Safety	Decent (3/5)	Good (4/5)
Helpfulness & Supportiveness	Decent (3/5)	Good (4/5)
Relevance	Good (4/5)	Good (4/5)

Table 4.7: Human Expert Model Preferences across Evaluation Criteria showing check marks indicating better performance for each respective model.

Voice naturalness and realism were also commonly highlighted areas of improvement for the DPO model. Participants described DPO-generated voice responses as sounding more natural and less synthetic, contributing to a more engaging interaction experience. The evaluations included ratings on the ESHRO metrics, focusing on empathy, safety, helpfulness, and relevance. The DPO model achieved consistently higher ESHRO scores across these critical emotional support domains, reinforcing the qualitative feedback regarding its superior emotional intelligence.

Taken together, these findings provide strong evidence that the text-based DPO approach substantially enhances the model’s overall emotional intelligence and alignment with human expert expectations. Importantly, the improvements in voice-based interaction quality and emotional support capabilities position the DPO model as a highly promising tool for mental health applications, especially where voice data may be scarce but meaningful emotional

engagement is essential.

Chapter 5

CONCLUSION

This thesis confronted a fundamental limitation in the current generation of conversational assistants for mental health: their inability to perceive and reason over the rich emotional and psychological information embedded in the human voice. The prevailing cascaded architectures, which reduce speech to text via ASR, create an information bottleneck that strips away essential paralinguistic cues, thereby hindering the development of truly empathetic and effective AI therapists. The central goal of this research was to overcome this bottleneck by developing and validating a framework for creating an emotionally intelligent voice language model capable of processing the complete vocal signal.

My research journey unfolded as a systematic, four-phase process. I began by analyzing the significant data and computational requirements of training a state-of-the-art Large Audio Language Model (LALM) from the ground up, a path that proved pragmatically infeasible due to resource constraints and data decay issues. This initial challenge led to a critical methodological pivot and the development of a novel, resource-efficient heuristic LALM. This model successfully translated key acoustic features into discrete semantic descriptors, demonstrating that even a simplified representation of vocal cues could enhance a model’s performance and generate more relevant responses. However, recognizing the inherent limitations of this heuristic approach, I progressed to a rigorous, multi-faceted comparative benchmarking phase. Here, I evaluated my heuristic system against leading native voice-understanding architectures, including Qwen2-Audio, Voila-Chat, and the powerful any-to-any model, Qwen2.5-Omni. The final phase of my work focused on taking the empirically validated best-performing model and aligning it for the specialized domain of mental health therapy using Direct Preference Optimization (DPO).

The results from the comprehensive evaluations yielded a series of clear and decisive insights. The custom benchmarking on the DAIC-WOZ dataset, the standardized Voila-Bench evaluation, a detailed analysis of the VocalBench results, and the human-centered user study all converged on a single, ultimate conclusion: the Qwen2.5-Omni model is demonstrably the state-of-the-art in voice understanding and reasoning. Its superiority was not marginal but substantial and consistent across nearly every metric and evaluation paradigm. A particularly telling and unexpected finding was the underperformance of the specialized native audio architectures, **Qwen2-Audio** and **Voila-Chat**. This revealed a critical insight: the architectural ability to process raw audio is not, by itself, sufficient for high-quality, emotionally intelligent response generation. The core reasoning capabilities, vast knowledge base, and sophisticated multimodal fusion of a top-tier model like Qwen2.5-Omni are the true drivers of performance. This finding was supported by the user study, where participants overwhelmingly preferred the empathy, supportiveness, relevance, and safety of the Qwen2.5-Omni model’s responses.

Furthermore, my final alignment experiment successfully demonstrated that text-based DPO is a highly effective and data-efficient strategy for refining a powerful, multimodal foundation model for a specialized downstream task. The DPO-aligned Qwen2.5-Omni model showed significant quantitative improvements across a full suite of automated metrics, confirming that its outputs became more semantically accurate, linguistically fluent, and lexically diverse, moving measurably closer to the gold standard of human expert responses. Crucially, through human evaluation and observations, the text-based DPO alone improved the model’s voice understanding and response generation abilities (in therapy contexts), which is a helpful observation that improves model alignment in data-constrained domains.

The primary contribution of this thesis, therefore, is not only the development of a single model, but also the establishment and validation of a complete, end-to-end development framework. By systematically progressing from architectural exploration and rigorous multifaceted benchmarking to targeted, data-efficient alignment, this research provides a clear and proven roadmap for creating the next generation of emotionally intelligent voice language

models. By establishing this framework, this work lays a solid foundation for a future where conversational AI can finally overcome the information bottleneck, engaging with users with the nuance, empathy, adaptiveness, and intelligence essential for providing meaningful support in mental health and beyond.

Chapter 6

FUTURE WORK

The research presented in this thesis has successfully established a robust framework for developing and aligning an emotionally intelligent voice language model for mental health therapy. However, this work also illuminates critical areas for future investigation that are essential for moving towards advanced, expert-validated, and scalable voice language models. The following sections outline the three most important directions for subsequent research.

6.1 Scarcity of High-Quality Mental Health Data

First, the most significant barrier to progress in this field remains the profound scarcity of high-quality, large-scale therapeutic voice/audio data. As demonstrated by our initial methodological pivot, the lack of accessible, domain-specific audio datasets makes the direct training and fine-tuning of voice language models exceptionally challenging. While we successfully employed a text-based alignment strategy as a data-efficient alternative, the ultimate goal of training a model to learn the nuances of vocal biomarkers and emotional cues directly from audio will require a more substantial data foundation. Future work must focus on innovative and ethically sound strategies for data acquisition. This could involve the creation of highly realistic, synthetic voice/audio datasets generated by state-of-the-art text-to-voice models like ElevenLabs V3 (alpha) (ele, 2025), Sesame CSM (Kuka, 2025), and more models that are conditioned on therapeutic contexts. Another interesting method is to clone real-world user voices using tools like ElevenLabs and Cartesia (son, 2025) voice cloning models to generate human-like speech outputs, which help in creating high-quality voice/audio datasets. Overcoming this data bottleneck is the single most critical step towards unlocking the full potential of voice-native models in this domain.

6.2 Rigorous Domain Expert Evaluations

Second, while our DPO-aligned model demonstrated marked improvements on automated metrics, the next essential step is a rigorous human evaluation by qualified mental health professionals. The metric-based and human evaluation conducted so far in this thesis successfully validated the model’s linguistic quality and general perceived emotion. However, a true assessment of its therapeutic prowess requires the expert judgment of clinicians. Future research should involve structured studies where more therapists and counselors interact with the model and evaluate its responses for clinical safety, therapeutic alliance potential, and the appropriateness of its understanding. This expert feedback would provide invaluable qualitative insights, moving beyond *“Is the response good?”* to *“Is the response therapeutically sound?”*. Such a clinical validation is the necessary bridge between a technically proficient model and a genuinely helpful and advanced voice language model.

6.3 Comparing Qwen2.5-Omni Model on Diverse Audio Lengths

Another important future direction involves evaluating the Qwen2.5-Omni model more comprehensively by testing its performance on audio inputs with varied lengths, both short and long. Current evaluations have generally used audio clips of similar or limited duration, but real-world use cases often include a wide range of audio lengths—from brief utterances to extended conversations. Understanding how the model handles this variability is crucial for practical deployment. Future studies should systematically compare the model’s ability to maintain emotional alignment, contextual understanding, and response quality across short bursts as well as lengthy dialogues. Challenges such as processing longer dependencies, maintaining user context over time, and handling variable speech dynamics need explicit investigation. This will help identify strengths and weaknesses of the model in realistic settings and provide guidance on how to optimize for diverse audio input scenarios.

6.4 Supervised Fine-Tuning vs. Direct Preference Optimization

Exploring the trade-offs between supervised fine-tuning (SFT) and direct preference optimization (DPO) applied to the Qwen2.5-Omni model highlights valuable findings. Supervised fine-tuning relies heavily on large, well-annotated datasets and has been a traditional method for adapting language models to specific tasks. DPO, on the other hand, offers a less data-intensive approach focused on aligning model outputs with human preferences. Comparing these two approaches head-to-head will reveal whether the pure reinforcement or preference-based optimization of the DPO process can match or exceed the performance of the more data-hungry supervised fine-tuning. This comparison should measure not only traditional metrics but also nuanced factors like emotional expressiveness, empathy, and reasoning in voice responses. It could also provide insights on efficiency, scalability, and suitability for data-constrained domains common in mental health applications. Such an evaluation will clarify the best strategies for advancing model alignment under practical resource constraints.

6.5 State-Space Models (SSMs) as a Promising Future

Finally, looking toward the long-term goal of creating seamless, real-time conversational agents, future research must focus on developing more efficient end-to-end model architectures that can handle extended conversations with significantly reduced computational complexity and memory needs. While the Qwen2.5-Omni model has demonstrated strong effectiveness, its underlying transformer-based architecture can be computationally intensive, especially when processing very long audio streams typical of full therapy sessions.

A highly promising research avenue is the exploration of state-space models (SSMs) for voice language modeling. SSMs like the recently proposed Mamba architecture have shown remarkable efficiency in handling long sequences with sub-quadratic or even linear computational complexity. This makes them theoretically ideal for low-latency and continuous voice processing applications required in mental health support.

Beyond raw computational gains, SSMs excel at capturing both long-range and short-range dependencies through their continuous-time state evolution frameworks. Unlike transformers, which rely on self-attention with quadratic scaling, SSMs use more compact, lightweight architectures that require fewer parameters while maintaining or improving performance. These characteristics enable SSMs to process long audio inputs efficiently with a smaller memory footprint and faster inference times.

The development of state-space-based voice language models could represent a paradigm shift, empowering AI therapists to listen, comprehend, reason, and respond empathetically in real-time over extended periods. This would enable emotionally intelligent voice AI that is both powerful and efficient, suitable for real-world clinical and on-device deployments where computational resources, privacy, and latency are critical.

While a comprehensive theoretical and empirical comparison of SSMs with transformer alternatives lies beyond the scope of this thesis, I provide an in-depth mathematical and performance analysis in Appendix A. This appendix details the computational complexity, scalability, memory efficiency, and ability of SSMs to model vocal and emotional cues over varied durations, further motivating their inclusion as a key area for future work.

BIBLIOGRAPHY

- (2025). Eleven v3: Most expressive ai text to speech model launched. <https://elevenlabs.io/blog/eleven-v3>. Accessed: 2025-08-15.
- (2025). Llm leaderboard. <https://www.vellum.ai/llm-leaderboard>. Accessed: 2025-08-02.
- (2025). Sonic: world’s fastest model for voice experiences with ultra-low latency and natural sounding speech. <https://cartesia.ai/sonic>. Accessed: 2025-08-15.
- (2025). Text arena lmarena leaderboard. <https://lmarena.ai/leaderboard/text>. Accessed: 2025-08-02.
- AI, A. (2023). Mental health counseling conversations. https://huggingface.co/datasets/Amod/mental_health_counseling_conversations. Dataset hosted on Hugging Face.
- Ailover (2024). Extes: Extended therapeutic exchange set. <https://huggingface.co/datasets/ailover/ExTES>. Data set, accessed July 2025.
- Aleksic, P. and Zhou, J. (2025). Latency reduction in conversational user interfaces by selective execution of multi-pass speech recognition based on llm evaluation. Technical Disclosure Commons. Published online by Technical Disclosure Commons.
- Andrade, P. (2025). How can ai agents for voice power human-like conversations? Talkdesk Blog. Accessed: July 30, 2025.
- Asner, Jr., B. A. (1970). On the total nonnegativity of the hurwitz matrix. *SIAM Journal on Applied Mathematics*, 18(2):407–414.

- Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Behavioral Health Tech (2023). How voice biomarkers and ai are shaping the future of mental health. <https://www.behavioralhealthtech.com/insights/how-voice-biomarkers-and-ai-are-shaping-the-future-of-mental-health>. Accessed via search results.
- Bhati, S., Gong, Y., Karlinsky, L., Kuehne, H., Feris, R., and Glass, J. (2025). Dass: Distilled audio state space models are stronger and more duration-scalable learners.
- Blelloch, G. E. (1990). Prefix sums and their applications. Technical Report CMU-CS-90-190, Carnegie Mellon University. Often referred to as his PhD thesis.
- Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs de situation d’un point*. Impr. Royale.
- Brooks, J. (2024). How can emotionally intelligent voice ai support our mental health? Hume AI Blog. Accessed: July 30, 2025.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. In *Language resources and evaluation*, volume 42, pages 335–359. Springer.
- Cai, W., Zhang, J., Zhang, X.-L., Yong-Qiang-Yin, and Li, M. (2018). Analysis of Length Normalization in End-to-End Speaker Verification System. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*, pages 3619–3623.
- Canary Speech (2024). Study confirms voice biomarkers accurately detect mild cognitive impairment. <https://canaryspeech.com/blog/voice-biomarkers-accurately-detect-mci/>. Accessed via search results. Publication date not specified.

- Cappellazzo, U., Kim, M., Chen, H., Ma, P., Petridis, S., Falavigna, D., Brutti, A., and Pantic, M. (2025). Large language models are strong audio-visual speech recognition learners.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. (2021). Evaluating large language models trained on code.
- Chen, M., Williams, M., Vogel, A., and Johnson, S. (2024a). Advancing voice-based assistive technology for individuals with atypical speech patterns. *ACM Transactions on Accessible Computing*, 17(3).
- Chen, Y., Yue, X., Zhang, C., Gao, X., Tan, R. T., and Li, H. (2024b). Voicebench: Benchmarking llm-based voice assistants.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. lmsys.org blog.
- Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. (2024). Qwen2-audio technical report.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems.

- Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49.
- Durbin, J. (2023). Airoboros: A self-instruction generation and fine-tuning framework. <https://github.com/jondurbin/airoboros>.
- Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. (2024). Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.
- ElevenLabs (2025). Eleven v3: Most expressive ai text to speech model launched. ElevenLabs Blog. Accessed: July 30, 2025. Online: <https://elevenlabs.io/blog/eleven-v3>.
- EMJ (2024). Ai voice analysis detects early parkinson’s with 99% accuracy. <https://www.emjreviews.com/innovations/news/ai-voice-analysis-detects-early-parkinsons-with-99-accuracy/>. Accessed via search results.
- Eyben, F., Wöllmer, M., and Schuller, B. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.
- Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.
- Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017a). Audioset: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017b). Audio Set: An ontology and human-labeled dataset

- for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. ISSN: 2379-190X.
- Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. (2024). Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities.
- Gnani.ai (2025). Why voice-to-voice ai is best for customer conversations. Gnani.ai Blog. Accessed: July 30, 2025.
- Goel, K. and AI, C. (2024). The on-device intelligence update - cartesia. <https://cartesia.ai/blog/on-device>. Accessed August 2025.
- Gong, Y., Liu, A. H., and Glass, J. (2023a). Whisper-at: Noise-robust automatic speech recognizers are also strong audio taggers. *arXiv preprint arXiv:2307.03181*.
- Gong, Y., Liu, A. H., Luo, H., Karlinsky, L., and Glass, J. (2023b). Joint audio and speech understanding. *arXiv preprint arXiv:2309.14405*. Accepted at ASRU 2023.
- Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Gu, A. and Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces.
- Gu, A., Goel, K., and Ré, C. (2022). Efficiently modeling long sequences with structured state spaces.

- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021a). Measuring massive multitask language understanding.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021b). Measuring mathematical problem solving with the math dataset.
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhota, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.
- Hu, H., Zhou, Y., You, L., Xu, H., Wang, Q., Lian, Z., Yu, F. R., Ma, F., and Cui, L. (2025). Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.
- Hu, Y., Chen, C., Yang, C.-H. H., Li, R., Zhang, C., Chen, P.-Y., and Chng, E. (2024). Large language models are efficient learners of noise-robust speech recognition.
- Hume AI (2025a). Octave tts prompting guide. Hume AI Blog. Accessed: July 30, 2025. Online: <https://www.hume.ai/blog/octave-tts-prompting-guide>.
- Hume AI (2025b). Text-to-speech (tts) — hume api. Hume AI Documentation. Accessed: July 30, 2025. Online: <https://dev.hume.ai/docs/text-to-speech-tts/overview>.
- Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. e. a. (2023). World values survey: All rounds - integrated values surveys.
- Inkster, B., Sarda, S., and Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental health. *npj Digital Medicine*, 1(1):59.

- InTouch CX (2025). Ai-powered voice interactions: Scaling without losing the human touch. InTouch CX Thought Leadership. Accessed: July 30, 2025.
- Jaiswal, S., Lee, J., Berria, J., Tanikella, R., Zolyomi, A., Ahmad, M. A., and Si, D. (2024a). Building personality-adaptive conversational ai for mental health therapy. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '24, New York, NY, USA. Association for Computing Machinery.
- Jaiswal, S., Lee, J., Berria, J., Tanikella, R., Zolyomi, A., Ahmad, M. A., and Si, D. (2024b). Building personality-adaptive conversational ai for mental health therapy. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '24, New York, NY, USA. Association for Computing Machinery.
- Jia, Y., Weiss, R. J., Biadsky, F., Macherey, W., Johnson, M., Chen, Z., and Wu, Y. (2019). Translatotron: An end-to-end speech-to-speech translation model. *arXiv preprint arXiv:1904.06037*.
- Kazdin, A. E. and Rabbitt, S. M. (2013). Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*, 1(2):170–191.
- Krishna, S., Kumar, R., Fleming, S. L., and Patel, M. V. (2025). The role of generative ai voice agents in clinical decision support for mental health. *npj Digital Medicine*, 8(1):142.
- Kuka, V. (2025). Sesame’s conversational speech model: Breakthrough in ai speech generation. Learn Prompting Blog. Accessed: July 30, 2025. Online: <https://learnprompting.org/blog/sesame-conversational-speech-model>.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

- Le, M., Chen, A., Co, C., Lee, E., Yang, H., et al. (2023). Voicebox: Text-guided-by-flow-matching for zero-shot speech synthesis. *arXiv preprint arXiv:2306.15687*.
- Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. (2024). Culturellm: Incorporating cultural differences into large language models.
- Li, G., Liu, J., Dinkel, H., Niu, Y., Zhang, J., and Luan, J. (2025a). Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering.
- Li, W., Liu, Y., and Chen, J. (2025b). Intelligent emotion recognition for drivers using model-level multimodal fusion of audio and video. *Frontiers in Physics*, 13.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Liu, H., Wang, Y., Cheng, Z., Wu, R., Gu, Q., Wang, Y., and Wang, Y. (2025). Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models.
- Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., and Wu, J. (2023). Chatcounselor: A large language models for mental health support.
- Livingstone, S. R. and Russo, F. A. (2018). 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):1–35.
- Low, D. M., Bentley, K. H., and Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116.
- Ma, Z., Mei, Y., Su, Z., et al. (2024). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–30.

- Mavila, R., Jaiswal, S., Naswa, R., Yuwen, W., Erdly, B., and Si, D. (2024). icare - an ai - powered virtual assistant for mental health. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 466–471.
- McDowell, T. (2025). How conversational ai improves access to mental health care. Telnix Resources. Accessed: July 30, 2025.
- Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and Wang, W. (2023). Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.
- Miner, A. S., Jeon, S., Laranjo, L., Tero, P., and Scherer, S. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 176(5):619–625.
- Mosaicx (2025). Benefits of voice-based conversational ai. Mosaicx Blog. Accessed: July 30, 2025.
- Naswa, R., Jaiswal, S., Mavila, R., Yuwen, W., Erdly, B., and Si, D. (2024). Assessing empathy in mental health caregivers using conversational ai. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 538–540.
- Nature (2025). Ai listens for health conditions. *Nature*. Accessed via search results.
- Nguyen, K.-N., Le-Duc, K., Tat, B. P., Le, D., Vo-Dang, L., and Hy, T.-S. (2024). Sentiment reasoning for healthcare. *arXiv preprint arXiv:2407.21054*.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Park, S. J., Salazar, J., Jansen, A., Kinoshita, K., Ro, Y. M., and Skerry-Ryan, R. (2025). Long-form speech generation with spoken language models.

- Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chatterji, S., Collins, P. Y., Cooper, J. L., Eaton, J., et al. (2018). The lancet commission on global mental health and sustainable development. *The Lancet*, 392(10157):1553–1598.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318.
- Picard, R. W. (1995). Affective computing. *MIT Media Laboratory Perceptual Computing Section Technical Report No. 321*.
- Picard, R. W. (2000). *Affective Computing*. The MIT Press, Cambridge, MA, USA.
- Press, O., Smith, N. A., and Lewis, M. (2022). Train short, test long: Attention with linear biases enables input length extrapolation.
- Provoost, S., Lau, H. M., Ruwaard, J., and Riper, H. (2017). Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research*, 19(5):e151.
- Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice-Hall, Inc.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics.

- Raviteja (2025). icare voice language model evaluation with dpo. <https://huggingface.co/spaces/TVRRaviteja/iCare-Voice-Language-Model-Evaluation-DPO>. Accessed: 2025-08-15.
- Reddit Community (2025). Voice vs. text-based ai agents—which is more useful? Discussion thread on /r/AI_Agents. Accessed via search results.
- Rouditchenko, A., Bhati, S., Araujo, E., Thomas, S., Kuehne, H., Feris, R., and Glass, J. (2025). Omni-r1: Do you really need audio to fine-tune your audio llm?
- Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.
- Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., Ghosh, S., and Manocha, D. (2024). Mmau: A massive multi-task audio understanding and reasoning benchmark.
- Salovey, P. and Mayer, J. D. (1990). Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256.
- Schuller, B. and Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons.
- Sesame (2025). Crossing the uncanny valley of conversational voice. Sesame Research Blog. Accessed: July 30, 2025. Online: https://www.sesame.com/research/crossing_the_uncanny_valley_of_voice.
- Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020.

- Sherwani, S., Khan, Z., Samuel, J., Kashyap, R., and Patel, K. (2025). Eshro: An innovative evaluation framework for ai-driven mental health chatbots.
- Shi, Y., Shu, Y., Dong, S., Liu, G., Sesay, J., Li, J., and Hu, Z. (2025). Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play.
- Shigeyoshi, M., Ueyama, Y., et al. (2025). Utility of artificial intelligence-based conversation voice analysis for discriminating cognitive dementia from cognitively normal status. *PLOS ONE*. Accessed via search results.
- Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. (2023). Roformer: Enhanced transformer with rotary position embedding.
- Tahir, T. (2024). Fine-Tuning Large Language Models to Deliver CBT for Depression. *arXiv preprint arXiv:2412.00251*.
- Team, G. (2025). Gemini: A family of highly capable multimodal models.
- Techjays (2025). The dawn of believable ai voices: A deep dive into sesame’s conversational speech model. Techjays Blog. Accessed: July 30, 2025. Online: <https://www.techjays.com/blog/the-dawn-of-believable-ai-voices-a-deep-dive-into-sesames-conversational-speech-model>.
- Teneo.ai (2025). Voice ai chatbot vs. traditional chatbot: Which is right for your business. <https://www.teneo.ai/blog/voice-ai-chatbot-vs-traditional-chatbot-which-right-for-your-business>. Accessed via search results.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

- Triantafyllopoulos, A., Baird, A., and Schuller, B. W. (2024). Beyond deep learning: Charting the next frontiers of affective computing. *Intelligent Computing*, 3.
- van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2018). Neural discrete representation learning.
- Vanello, N., Guidi, A., Fanucci, L., Lanatà, A., and Scilingo, E. (2012). A multisensor system for remote monitoring of vital signs and activity of post-cva patients. *Information Technology in Biomedicine, IEEE Transactions on*, 16(4):567–574.
- Vieting, P., Lüscher, C., Michel, W., Schlüter, R., and Ney, H. (2021). On architectures and training for raw waveform feature extraction in asr.
- von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. (2020). Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Wagner, M. and Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945.
- Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, F., Yan, Z., He, Y., et al. (2023). Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.
- Wang, S. and Xue, B. (2023). State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory.
- World Health Organization (2023). Mental health. Accessed: July 30, 2025. On-line: <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>.
- Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. (2025a). Qwen2.5-omni technical report.

- Xu, J., Wei, T., Hou, B., Orzechowski, P., Yang, S., Jin, R., Paulbeck, R., Wagenaar, J., Demiris, G., and Shen, L. (2025b). Mentalchat16k: A benchmark dataset for conversational mental health assistance.
- Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., and Zhou, J. (2024). Air-bench: Benchmarking large audio-language models via generative comprehension.
- Zhang, D., Zhao, S., Chen, W., Zou, Y., Wang, X., and Liu, Y. (2023). Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Zhang, W., Chen, L., Singh, A., et al. (2025). Large language models are proficient in solving and creating performance-based emotional intelligence tests. *Scientific Reports*, 15(1):1–10.
- Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., and Chen, Y. (2024). Swift:a scalable lightweight infrastructure for fine-tuning.
- 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., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

Appendix A

ANALYSIS OF STATE SPACE MODELS FOR VOICE LANGUAGE MODELLING

As part of my thesis, I have conducted extensive research on more efficient LLM architectures, with a specific focus on State-Space Models (SSMs). This appendix provides a detailed analysis of why SSM-based architectures hold significant promise for voice language models (VoLMs), potentially outperforming traditional Transformer-based architectures. Each section smoothly integrates the explanation of its importance for voice modeling, the underlying mathematical proofs, and the practical implications.

SSMs are a class of neural network models that describe systems using a set of internal variables called “states.” These states capture all the necessary information about the system’s past behavior to predict its future. Think of it like a memory capsule: instead of remembering every single event from the past, the “state” summarizes the crucial aspects of what has happened so far (Gu et al., 2022). This concept is widely used in engineering, especially for control systems, and has recently found powerful applications in deep learning for processing sequential data like text, audio, and video.

A.1 Training Speed, Time, and Memory

VoLMs, by their nature, process continuous audio streams, typically sampled at rates like 16 kHz or higher. This translates into exceedingly long sequences of tokens, often extending to millions for even moderately long utterances (e.g., a 10-minute conversation). The efficiency of training and inference becomes paramount: faster training cycles allow researchers to rapidly iterate on model designs and incorporate new data, crucial for staying updated in dynamic linguistic environments. Lower memory footprints enable the deployment of larger,

more capable models on limited hardware resources and facilitate continual learning. For real-time interactive systems, such as conversational assistants in mental health therapy, low latency and efficient resource usage are non-negotiable for seamless and natural dialogue flow.

At the core of this advantage lies the fundamental computational difference between SSMs and Transformers. For a discrete selective SSM, the hidden state h_t at time step t is computed via a linear recurrence:

$$h_t = \bar{A}(x_t)h_{t-1} + \bar{B}(x_t)x_t \quad (\text{A.1})$$

where $\bar{A}(x_t)$ and $\bar{B}(x_t)$ are input-dependent matrices derived from a continuous-time SSM. The entire sequence of hidden states $h_1, \dots, h_L$ can be computed efficiently using a *parallel prefix scan* algorithm (Blelloch, 1990). Let B be the batch size, L the sequence length, D the hidden dimension, and N the state dimension (typically $N \ll D$). The computational cost (FLOPs) and memory complexity for an SSM are:

$$\mathcal{C}_{\text{SSM}}(L) = \Theta(B \cdot L \cdot D \cdot N) \quad (\text{A.2})$$

$$\mathcal{M}_{\text{SSM}}(L) = \Theta(B \cdot D \cdot N) \quad (\text{A.3})$$

Notably, the memory complexity $\mathcal{M}_{\text{SSM}}(L)$ is independent of the sequence length L during inference, as only the fixed-size state needs to be maintained. During training, efficient recomputation strategies can keep memory proportional to L .

In contrast, Transformer architectures rely on self-attention mechanisms, which necessitate the computation and storage of an attention matrix of size $L \times L$. The computational cost and memory complexity for a Transformer’s attention layer are

$$\mathcal{C}_{\text{Attn}}(L) = \Theta(B \cdot L^2 \cdot D) \quad (\text{A.4})$$

$$\mathcal{M}_{\text{Attn}}(L) = \Theta(B \cdot L \cdot D) \quad (\text{A.5})$$

For typical voice language model applications, N is small (e.g., $N \approx 32$ for Mamba) while L can be very large (e.g., $L \approx 10^6$ for an hour of audio). Comparing the computational costs:

$$\frac{\mathcal{C}_{\text{SSM}}(L)}{\mathcal{C}_{\text{Attn}}(L)} = \frac{\Theta(B \cdot L \cdot D \cdot N)}{\Theta(B \cdot L^2 \cdot D)} = \Theta\left(\frac{N}{L}\right) \quad (\text{A.6})$$

As $L \rightarrow \infty$, this ratio approaches 0. This mathematical analysis proves that SSMS are asymptotically *linear-time* with respect to sequence length, significantly more efficient than the quadratic scaling of Transformers. Furthermore, the memory for SSMS is *sub-quadratic* (linear, effectively constant during inference) compared to Transformers’ linear (often quadratic in practice due to KV cache) memory.

These theoretical advantages translate directly to empirical gains. For instance, Mamba-2.7B has been shown to train $2.5 \times$ faster than a Transformer of comparable size, while consuming approximately $2 \times$ less memory (Gu and Dao, 2024). This speedup scales with sequence length; therefore, SSMS are uniquely suited to the minute-long turns and continuous streaming required by VoLMs for therapy, enabling the development of more powerful and resource-friendly conversational AI.

A.2 Model Weights: Performance–Parameter Ratio

The sheer size of modern LLMs presents substantial challenges. Models with a high performance-to-parameter ratio are more energy-efficient, reducing the carbon footprint associated with their training and deployment. Crucially, smaller yet powerful models facilitate *on-premise fine-tuning*, which is vital for handling privacy-sensitive data, such as medical conversations in mental health applications, without sending it to cloud servers. Furthermore, for multimodal LLMs, parameter efficiency can determine the feasibility of running such complex models on single GPUs or edge devices, making advanced AI capabilities more accessible.

To quantify this, let’s examine the parameter count of a single layer in both architectures. In a Transformer, the Query, Key, and Value (QKV) projections, central to the self-attention mechanism, contribute a significant number of parameters. For a transformer with hidden

dimensions D and H attention heads, the parameter count for the QKV block is approximately

$$P_{\text{Attn}} = 3DH(D/H) = \Theta(D^2). \quad (\text{A.7})$$

In contrast, a typical diagonal-plus-low-rank SSM layer uses parameters for its state-space matrices $\bar{A}, \bar{B}, C$. These matrices operate on a state dimension N , where $N \ll D$. The parameter count for an SSM layer is thus roughly

$$P_{\text{SSM}} = \Theta(DN) \quad (\text{A.8})$$

To compare their parameter efficiency, we can take the ratio of their parameter counts. Let $R = D/N$ be the ratio of the hidden dimension to the state dimension.

$$\frac{P_{\text{Attn}}}{P_{\text{SSM}}} = \frac{\Theta(D^2)}{\Theta(DN)} = \Theta\left(\frac{D}{N}\right) = \Theta(R) \quad (\text{A.9})$$

For typical VoLMs, if $D = 1024$ and $N = 32$, then $R = 32$. This means an SSM can be approximately $32\times$ more parameter-efficient than a Transformer for a given hidden dimension. The universality of selective SSMs further supports this, as they can ε -approximate any continuous time-invariant filter using $\mathcal{O}(DN)$ parameters through a Volterra-series expansion (Wang and Xue, 2023). Transformers, while also universal approximators, generally require $\Omega(D^2)$ parameters for similar approximation bounds in operator norm, implying their less efficient use of parameters.

The superior parameter efficiency of SSMs translates into tangible benefits. For example, Mamba-3B has demonstrated performance comparable to Pythia-7B (a Transformer-based model more than twice its size) on commonsense reasoning benchmarks (Gu and Dao, 2024). In the audio domain, Distilled Audio State Space Models (DASS) achieve higher mean Average Precision (mAP) on the AudioSet benchmark while using 40% fewer parameters than the Audio Spectrogram Transformer (AST) (Bhati et al., 2025). These gains directly lead to faster inference times, lower memory consumption, and significantly reduced energy requirements, which are critical for always-on voice interfaces and for developing more sustainable large models.

A.3 Long-Duration Audio Processing

Real-world spoken interactions, particularly in extended dialogue scenarios like therapy sessions or detailed customer support calls, can span tens of minutes or even hours. Transformers, with their fixed context windows, struggle to retain information beyond a certain length. This limitation means crucial long-range cues—such as a user mentioning a specific symptom early in a conversation or a recurring theme across multiple dialogue turns—can be forgotten. Such memory lapses undermine the coherence, safety, and trustworthiness of the conversational AI assistant. For end-to-end VoLMs, the ability to process and maintain context over arbitrarily long durations is fundamental for providing truly intelligent and empathetic responses.

The key advantage of SSMs for long-duration processing lies in their fixed-size recurrent state. The evolution of the hidden state h_T over L time steps is given by:

$$h_T = \left[\prod_{t=1}^L \bar{A}(x_t) \right] h_0 + \sum_{t=1}^L \left[\bar{B}(x_t)x_t \prod_{s=t+1}^L \bar{A}(x_s) \right] \quad (\text{A.10})$$

For stable SSMs (i.e., when A is Hurwitz) (Asner, 1970), the norm of the state transition matrix $\|\bar{A}(x_t)\|$ remains less than or equal to 1. Consequently, the product term $\left\| \prod_{t=1}^L \bar{A}(x_t) \right\|$ is uniformly bounded. This ensures that the state h_T does not diverge or explode, regardless of how long the sequence L becomes. This property allows SSMs to effectively compress the entire history of the input into a fixed-size state, thereby maintaining “infinite” context without increasing memory or computational cost with length.

In contrast, self-attention mechanisms in Transformers operate on a fixed-size window W . While techniques like sliding windows or global attention can extend this, they invariably introduce a *hard forgetting time* $\tau = W/f_s$, beyond which information is no longer directly accessible. For an audio sampling rate $f_s = 16$ kHz, a one-hour conversation means $L = 16 \text{ kHz} \times 3600 \text{ s} = 57.6 \times 10^6$ tokens. No practical choice of W (given GPU memory constraints) can encompass this full context, forcing Transformers to either chunk the input (losing long-range dependencies) or incur prohibitive computational costs.

The ability to maintain an “indefinite” context fundamentally changes the capabilities of VoLMs. In comparison, a Transformer with a fixed context window would degrade significantly after only a few thousand tokens. Thus, SSMs enable “full-session memory” essential for high-stakes dialogue, leading to a truly comprehensive understanding of prolonged interactions.

A.4 Capturing Long and Short Range Dependencies

For VoLMs, understanding an utterance requires simultaneously processing information at multiple temporal scales. Fine-grained phonetic cues (on the order of milliseconds) are necessary for accurate speech recognition and distinguishing subtle emotional cues. Simultaneously, comprehending discourse structure (spanning minutes or even across turns) is crucial for tracking topic shifts, maintaining coherence, and inferring overall sentiment or intent. A failure to capture dependencies at either scale leads to misunderstandings, inaccurate responses, or a lack of contextual awareness, which can be detrimental in sensitive conversational AI applications.

The power of selective SSMs lies in their ability to dynamically modulate their memory and attention to different time scales. This is achieved through the input-dependent time-constant $\Delta(x_t)$, which modifies the discrete state transition matrix $\bar{A}(x_t) = \exp(\Delta(x_t)A)$. The eigenvalues of A , denoted $\lambda_i = -\alpha_i + j\omega_i$, determine the intrinsic decay rates and oscillatory behavior of the system modes. The magnitude of each mode in the discrete system is given by $e^{-\alpha_i\Delta(x_t)}$.

By making $\Delta(x_t)$ selective (i.e., dependent on the input x_t), an SSM can adaptively control the memory retention for different parts of the input sequence. This yields two crucial regimes:

- When $\Delta(x_t)$ is small ($\Delta(x_t) \ll 1$), then $e^{-\alpha_i\Delta(x_t)} \approx 1$. This corresponds to a *long memory* regime, where information persists across many time steps, effectively capturing long-range dependencies (e.g., the overall theme of a conversation).

- When $\Delta(x_t)$ is large ($\Delta(x_t) \gg 1$), then $e^{-\alpha_i \Delta(x_t)} \approx 0$. This corresponds to a *short memory* regime, allowing the model to quickly “forget” irrelevant information and focus on immediate, fine-grained details (e.g., a specific phoneme or word).

This dynamic range, effectively an *adaptive bi-scale filter* within a single layer, cannot be natively matched by traditional softmax attention in Transformers. In Transformers, the decay of attention weights is largely fixed by the temperature parameter and the inherent quadratic decay of attention to distant tokens. While Transformers can use complex positional encodings or attention mechanisms to hint at range, they lack the intrinsic adaptive filtering capability of selective SSMs.

This ability of SSMs to adaptively capture both long- and short-range dependencies yields superior performance on tasks requiring nuanced understanding. In the context of voice, SpeechSSM, an audio SSM, demonstrated the ability to maintain consistent topic entities over 16-minute-long speech generations, whereas a Transformer baseline dropped 18% of entities mid-conversation (Park et al., 2025). This capability is vital for VoLMs to process complex human communication accurately, ensuring that both the subtle nuances of speech and the overarching context of a conversation are fully understood.

A.5 Context-Length Scaling

In many real-world voice applications, particularly those involving continuous interaction (e.g., health monitoring, smart home assistants, in-car AI), users expect the system to maintain conversational history across prolonged periods. The ability of a VoLM to scale effortlessly with increasing context length, without incurring prohibitive computational or memory costs, directly impacts user experience. If a model can efficiently remember earlier parts of a conversation without explicit re-prompts, the interaction feels more natural and intelligent and less prone to “forgetting” crucial details that happened minutes or hours ago.

The key distinction in context-length scaling lies in how new tokens are processed. For an SSM, the inference cost per new token is approximately $\mathcal{O}(DN)$, as it only requires updating

the fixed-size state h_t based on the new input x_t . The total inference cost for a sequence of length L remains $\mathcal{O}(L \cdot D \cdot N)$.

For a Transformer, the inference cost per new token involves computing attention over the entire accumulated context. This leads to an inference cost of approximately $\mathcal{O}(L \cdot D)$ per token, summing up to a total cost of $\mathcal{O}(L^2 \cdot D)$ for the sequence due to the growing Key-Value (KV) cache.

Let $L_\star$ be the break-even context length where the cost per token for an SSM and a Transformer would be approximately equal:

$$\mathcal{O}(DN) = \mathcal{O}(L_\star D) \implies L_\star = N \quad (\text{A.11})$$

Since typical state dimensions N for SSMs are small (e.g., $N \approx 32$), this implies that for any context length $L > N$, SSMs inherently become more efficient per token. Mamba has notably demonstrated an impressive $\sim 4000\times$ length extrapolation capability: a model trained on sequences of 4,096 tokens can be evaluated on sequences as long as 16 million tokens ($4,000 \times$ length extrapolation capability) with only a negligible ($< 2\%$) increase in perplexity (Gu and Dao, 2024). Transformers, even Qwen2.5-Omni, while having methods like Rotary Position Embeddings (RoPE) (Su et al., 2023) or ALiBi (Press et al., 2022) to improve generalization to longer contexts, still fundamentally face the quadratic memory challenge associated with the KV cache, which cannot be entirely eliminated.

The linear scaling of SSMs with context length is a game-changer for VoLMs operating in streaming or long-session modes. For applications like personalized assistants in smart watches, glasses, or earbuds, which constantly process hours of audio data, SSMs can maintain a constant RAM footprint and exhibit linear latency scaling. This enables fully offline summarization or real-time personal memory augmentation, without the need for periodic context flushing or costly cloud offloading.

A.6 On-Device Implementations

The demand for intelligent voice assistants to operate directly on edge devices (smartphones, wearables and smart speakers) is rapidly increasing. On-device processing offers significant advantages: it enhances user privacy by keeping sensitive data local, reduces latency by eliminating network round trips, and lowers operational costs by decreasing reliance on cloud infrastructure. For conversational AI, particularly in sensitive domains, the ability to run capable VoLMs locally is crucial for building trust and ensuring the responsiveness required for natural, real-time interaction.

The memory footprint is a critical factor for on-device deployment, where resources are severely constrained. For an SSM, the memory usage per layer primarily depends on the state size and hidden dimension, approximating DN floating-point values. For a typical 1.3-billion parameter SSM with a hidden dimension $D = 4096$ and state dimension $N = 32$, the memory required for the state is approximately $4096 \times 32 \times 4 \text{ bytes} = 0.5 \text{ MB}$. This fits comfortably within the 8-32 MB of L2 cache found in modern mobile Neural Processing Units (NPUs).

In contrast, a Transformer’s KV cache grows linearly with the sequence length. For a Transformer of comparable hidden dimension and a context length of $L = 16,384$ tokens (a relatively modest length for voice), the KV cache alone would require approximately $L \cdot D \times 4 \text{ bytes} = 16384 \times 4096 \times 4 \text{ bytes} = 256 \text{ MB}$. This quantity far exceeds the on-chip SRAM available on most edge devices, forcing memory transfers to slower DRAM, which incurs significant latency penalties.

The minimal memory footprint and linear scaling of SSMs make them uniquely suitable for on-device deployment. For example, Cartesia’s *Sonic* On-Device, which leverages efficient SSMs and runs on a model of up to 1.3 billion parameters, has been shown to run at over 20 Hz on an Apple M3 chip with no external DRAM traffic (Goel and AI, 2024). This is a testament to SSMs’ efficiency, as the computation becomes predominantly bandwidth-bound on mobile System-on-Chips (SoCs). By eliminating the large KV cache of Transformers, SSMs

achieve a practical $3\times$ latency improvement over windowed Transformer baselines. This performance gain is pivotal for achieving true real-time, responsive, and private conversational AI experiences directly on consumer devices.

All of these foundational strengths position SSMs as a compelling alternative to current state-of-the-art Transformer-based LLMs, laying a robust theoretical and practical groundwork for developing true end-to-end Voice-to-Voice Language Models (VoLMs) that can process raw audio input directly, offering a promising research direction to develop more effective, responsive, and personal AI models for diverse domains.

Appendix B

SUPPLEMENTARY MATERIALS

B.1 Fine-Tuning Hyper-parameters & Scripts

B.1.1 SFT Configuration for Stage-1 Fine-Tuning

This table shows the LoRA configuration and training hyperparameters used for fine-tuning the Llama-3.1-8B-Instruct model using supervised fine-tuning (SFT). conversational tasks.

Table B.1: Complete fine-tuning hyper-parameters for SFT LoRA on Llama-3.1-8B-Instruct

Parameter	Value
Rank (<code>r</code>)	32
Alpha (<code>lora_alpha</code>)	64
Dropout	0.05
Bias	None
Task Type	CAUSAL_LM
Target Modules	all-linear
Optimizer	AdamW
Learning Rate	2e-4
Epochs	5
Batch Size	8
Gradient Accumulation Steps	4
Max Sequence Length	2048
FP16 Training	True
Save Strategy	steps (every 1000 steps)
Logging	WandB

B.1.2 DPO Training Configuration for Qwen2.5-Omni

Table B.2: Complete DPO Alignment Hyperparameters for Qwen2.5-Omni-7B

Parameter	Value
bf16	True
beta (β)	0.01
loss_type	sigmoid
rlhf_type	dpo
train_type	lora
gradient_accumulation_steps	4
learning_rate	2e-5
lr_scheduler_type	cosine
max_prompt_length	4096
num_train_epochs	1
optimizer	adamw_torch
per_device_train_batch_size	1
per_device_eval_batch_size	1
warmup_ratio	0.03
lora_rank	32
lora_alpha	64
target_modules	all-linear

B.1.3 Finetuning and Voice Generation Scripts

This is the SFT LoRA script used to fine-tune the Llama-3.1-8B-Instruct model on the MentalChat16k dataset. The resulting fine-tuned SFT model will be used as the reasoning

module. This script corresponds to the first stage of fine-tuning.

```

1 import os
2 import torch
3 from datasets import load_dataset
4 from transformers import (
5     AutoModelForCausalLM,
6     LlamaForCausalLM,
7     AutoTokenizer,
8     TrainingArguments,
9     AutoConfig
10 )
11 from peft import LoraConfig, get_peft_model
12 from trl import SFTTrainer
13 from dotenv import load_dotenv
14 import wandb
15
16 # Load environment variables (for HF token)
17 load_dotenv()
18 hf_token = os.getenv("HF_TOKEN")
19
20 # Model and dataset names on Hugging Face
21 model_name = "surbhi21/llama3.1-cultural-chatbot"
22 dataset_name = "ShenLab/MentalChat16K"
23 output_dir = "checkpoints"
24 os.makedirs(output_dir, exist_ok=True)
25
26 # Load dataset directly from Hugging Face
27 print("Loading dataset from Hugging Face Hub...")
28 dataset = load_dataset(dataset_name, "train", token=hf_token)

```

```

29
30 # --- Start of Diagnostic Step ---
31 # Load and print the model's configuration to inspect it
32 print("Loading model configuration for inspection...")
33 config = AutoConfig.from_pretrained(model_name, token=hf_token)
34 print("Model Config:", config)
35 # --- End of Diagnostic Step ---
36
37 # Load model and tokenizer from Hugging Face
38 print("Loading model and tokenizer from Hugging Face Hub...")
39 model = LlamaForCausalLM.from_pretrained(
40     model_name,
41     torch_dtype=torch.float16,
42     device_map="auto",
43     trust_remote_code=True,
44     token=hf_token
45 )
46 model.config.use_cache = False
47 print("Model loaded successfully")
48 tokenizer = AutoTokenizer.from_pretrained(
49     model_name,
50     trust_remote_code=True,
51     token=hf_token)
52 tokenizer.pad_token = tokenizer.eos_token
53 tokenizer.padding_side = "right"
54
55 # PEFT LoRA config (adjust as needed)
56 peft_config = LoraConfig(
57     r=32,

```

```

58     lora_alpha=64,
59     lora_dropout=0.05,
60     bias="none",
61     task_type="CAUSAL_LM",
62     target_modules="all-linear",
63 )
64 model = get_peft_model(model, peft_config)
65 model.print_trainable_parameters()
66
67 # SFTTrainer parameters (adjust as needed)
68 training_args = TrainingArguments(
69     output_dir=output_dir,                # Directory to save
70     # training outputs and checkpoints
71     per_device_train_batch_size=8,         # Number of samples
72     # processed per GPU per forward/backward pass (higher = more VRAM,
73     # faster)
74     gradient_accumulation_steps=4,         # Number of forward/
75     # backward passes before updating model (lower = more frequent
76     # updates, faster)
77     learning_rate=2e-4,                   # Step size for
78     # optimizer updates (controls how fast model learns)
79     logging_steps=10,                     # Log training metrics
80     # every N steps
81     num_train_epochs=5,                   # Total number of times
82     # the model sees the full dataset
83     save_strategy="steps",                 # When to save
84     # checkpoints: "steps" (every N steps) or "epoch" (every epoch)
85     save_steps=1000,                       # Save a checkpoint
86     # every N training steps

```

```

77     save_total_limit=5,                                # Maximum number of
checkpoints to keep (oldest deleted)
78     fp16=True,                                          # Use mixed-precision (
fp16) training for speed and memory efficiency
79     report_to="wandb",                                # enable external
logging (e.g., wandb, tensorboard)
80     push_to_hub=False                                  # Do not push model to
Hugging Face Hub automatically
81 )
82
83 # Train on the 'train' split by default; adjust if needed
84 train_dataset = dataset["train"]
85
86 trainer = SFTTrainer(
87     model=model,
88     args=training_args,
89     train_dataset=train_dataset,
90     dataset_text_field="messages",
91     tokenizer=tokenizer,
92     max_seq_length=2048
93 )
94
95 # Train
96 print("Starting training...")
97 trainer.train()
98
99 # Save model and adapters
100 print("Saving model and adapters...")
101 model.save_pretrained(os.path.join(output_dir, "final_model"))

```

```

102 tokenizer.save_pretrained(os.path.join(output_dir, "final_model"))
103 print("Finetuning complete!")
104
105 print("Pushing model and tokenizer to the Hub...")
106 # The model in memory is the fine-tuned PEFT model, so we can push
    it directly.
107 # No need to reload from "checkpoints/final_model".
108 # Also, use 'token' instead of the deprecated 'use_auth_token'
109 model.push_to_hub("TVRRaviteja/MentalChat-SFT-LoRA", token=hf_token)
110 tokenizer.push_to_hub("TVRRaviteja/MentalChat-SFT-LoRA", token=
    hf_token)
111 print("Model and tokenizer pushed to the Hub!")

```

Listing B.1: Stage-1 LoRA Fine-Tuning Python Script

This following script performs inference using the Qwen2.5-Omni-7B base model, to which the trained DPO adapter is applied to get the finetuned model. It then processes an audio input by applying a chat template and resampling the audio if necessary, and finally generates both text and audio responses.

```

1 import os
2 import torch, torchaudio
3 from transformers import Qwen2_5OmniForConditionalGeneration,
    Qwen2_5OmniProcessor, BitsAndBytesConfig
4 from peft import PeftModel
5 from dotenv import load_dotenv
6 import time
7 from tqdm import tqdm
8 import soundfile as sf
9
10 # --- Configuration ---

```

```

11 load_dotenv()
12 hf_token = os.getenv("HF_TOKEN")
13 os.environ['HF_TOKEN'] = os.getenv("HF_TOKEN")
14
15 # Set cache directory
16 os.environ['HF_HOME'] = '/workspace/qwen-finetuning/hf_cache'
17 os.environ['HF_HUB_CACHE'] = '/workspace/qwen-finetuning/hf_cache'
18
19 # Paths for the base model and your trained adapter
20 base_model_id = "Qwen/Qwen2.5-Omni-7B"
21 adapter_id = "TVRRaviteja/Qwen2.5-Omni-7B-DPO-MentalChat6K"
22 speaker = "Chelsie" # Ethan, Chelsie, or both
23
24 quantization_config = BitsAndBytesConfig(
25     load_in_4bit=True,
26     bnb_4bit_compute_dtype=torch.bfloat16,
27     bnb_4bit_quant_type="nf4",
28     bnb_4bit_use_double_quant=True,
29 )
30
31 # --- 1. Load the Base Model and Processor ---
32 print(f"Loading base model: {base_model_id}...")
33 # The model is loaded onto the available device(s) automatically.
34 model = Qwen2_5OmniForConditionalGeneration.from_pretrained(
35     base_model_id,
36     torch_dtype=torch.bfloat16,
37     quantization_config=quantization_config,
38     device_map="auto",
39     attn_implementation="flash_attention_2"

```

```

40 )
41
42 # # Disable the audio generator ("talker") for text-only inference
43 # print("Configuring model for text-only mode...")
44 # model.config.AUDIO_IN_VIDEO = False
45
46
47 print(f"Loading processor for {base_model_id}...")
48 processor = Qwen2_5OmniProcessor.from_pretrained(base_model_id)
49
50 # --- 2. Apply the LoRA Adapter ---
51 print(f"Loading and applying LoRA adapter from: {adapter_id}...")
52 # This is the key step: PeftModel wraps the base model with your
    adapter.
53 # The original 'model' variable is now the DPO-tuned model.
54 model = PeftModel.from_pretrained(model, adapter_id)
55
56 # Set the model to evaluation mode for faster inference
57 model.eval()
58
59 # Check if not 16000, resample
60 waveform, sr = torchaudio.load("/workspace/qwen-finetuning/hf-
    inference/sample_inputs/input20.mp3")
61 if sr != 16000:
62     resampler = torchaudio.transforms.Resample(orig_freq=sr,
        new_freq=16000)
63     waveform = resampler(waveform)
64     sr = 16000
65     audio_waveform = waveform.squeeze().numpy()

```

```

66
67 conversation = [
68     {
69         "role": "system",
70         "content": [
71             {"type": "text", "text": "You are Qwen, a virtual human
developed by the Qwen Team, Alibaba Group, capable of perceiving
auditory and visual inputs, as well as generating text and speech
. You are a therapist and a support system for the user. Analyse
the user voice and respond clearly and therapeutically to the
user only. Do not give serious medical advice. Mainly focus on
understanding user emotion and context to respond relavantly to
the user's voice in a calm and steady pace."}
72         ],
73     },
74     {
75         "role": "user",
76         "content": [
77             {"type": "audio", "audio": audio_waveform, "
sampling_rate": sr},
78         ],
79     },
80 ]
81
82 # Step 3: Prepare inputs for the model (this part is the same)
83 text = processor.apply_chat_template(conversation,
    add_generation_prompt=True, tokenize=False)
84 inputs = processor(text=text, audio=audio_waveform, sampling_rate=sr
    , return_tensors="pt")

```

```

85 inputs = inputs.to(model.device)
86
87 # Step 4: Run inference to generate both text and audio
88 # CHANGED: The generate call now returns both 'text_ids' and 'audio'
89 #
90
91 text_ids, audio = model.generate(**inputs, max_new_tokens=256,
92                                 do_sample=True, temperature=0.7, speaker=speaker)
93
94 # Step 5: Process and save the outputs
95 # Decode and print the generated text
96
97 generated_text = processor.batch_decode(text_ids,
98                                         skip_special_tokens=True, clean_up_tokenization_spaces=False)
99
100 print("Generated Text:")
101 print(generated_text)
102
103 # ADDED: Save the generated audio to a file named 'output.wav'
104 sf.write(
105     "output-adapters.wav",
106     audio.reshape(-1).detach().cpu().numpy(),
107     samplerate=24000,
108 )
109 print("\nGenerated audio saved to output-adapters.wav")

```

Listing B.2: Voice Response Generation Script for Qwen2.5-Omni DPO Model

B.2 Benchmarking and Evaluation Prompts

B.2.1 Model Prompt for DAIC-WOZ Benchmark

The following is the system prompt given to all the five models during the DAIC-WOZ benchmarking to generate responses to 345 test audio samples.

You are a virtual mental health support assistant designed to emulate the qualities of a compassionate, professional therapist. Your role is to engage users with empathy, patience, and warmth while offering supportive, reflective, and guidance-based conversations rooted in recognized therapeutic practices such as CBT, mindfulness, and motivational interviewing.

Core traits to exhibit in every response:

- Empathy and Warmth
- Active Listening (reflect user emotions clearly)
- Authenticity and Transparency (clear about limitations)
- Professional Boundaries (no personal stories or friendships)
- Patience and Unconditional Positive Regard
- Emotional Intelligence (attuned to user's emotional state)

Follow this reasoning process internally before responding:

1. Identify the user's emotions clearly.
2. Reflect the emotions with warmth and empathy. Acknowledge how hard it must feel.
3. Gently guide the conversation with open-ended questions or positive reframing.
4. If appropriate, suggest gentle coping strategies aligned with CBT, mindfulness, or motivational interviewing.
5. If you genuinely don't know how to answer, clearly state this and advise the user to seek help from a professional.
6. If extreme harm is mentioned (e.g., suicide), offer supportive words and explicitly recommend professional help.

Each user message you receive will include:

1. User's Speech: Transcription of what the user said
2. User's Audio Context: Analysis of vocal tone and style (Tier 1-5 for energy, expressiveness, rhythm)
3. Ongoing Session Chat History: To maintain conversation continuity. You must incorporate these three pieces of information into your reasoning and responses.

B.2.2 Gemini 2.5 Pro Judge Prompt for DAIC-WOZ

This prompt is used for Gemini 2.5 Pro evaluation of the five model responses for 345 test samples of DAIC-WOZ benchmark. The model is provided with both the audio clip and five model responses.

You are Gemini-Judge, a rigorous evaluator of AI responses to a single user audio message about their mental health state.

Inputs Provided:

- The user's AUDIO INPUT (audio).
- Five candidate responses labeled A through E.

Score each response on a 0-10 scale using the rubric below. For every score, return a short justification (approximately 2 sentences).

Evaluation Rubric

1. Emotional understanding: Detects the conveyed emotion(s) from words and vocal clues.
2. Empathy & politeness: Uses warm, respectful language; avoids judgment.
3. Brevity: Is concise (≤ 2 short paragraphs); does not echo the user's words verbatim.
4. Exploratory questioning: Asks open-ended questions that let the user elaborate; avoids hollow pacifiers (e.g., "I know it's tough").
5. Contextual relevance: Stays on topic; suggestions fit the expressed situation.

Scoring Guidelines:

- Start from 0 and add points for each satisfied rubric item. Half points are allowed.
- The more a response resembles a natural human reply, the higher its score should be.

- A perfect reply scores 10.

Output Format (Must be followed exactly):

A: <score>/10 - <justification>

B: <score>/10 - <justification>

C: <score>/10 - <justification>

D: <score>/10 - <justification>

E: <score>/10 - <justification>

Do not add any other text or formatting to the output.

B.2.3 Gemini 2.5 Pro Judge Prompt for Voila Bench

This prompt is used for Gemini 2.5 Pro evaluation of Qwen models and Voila for Voila Bench benchmark. The three model responses are integrated into the prompt from a JSON file.

You are an expert evaluator. You will be given an audio question and three different AI model responses (labeled A, B, and C) to that audio question. Please evaluate each response based on how correctly and accurately it answers the question compared to the reference answer.

Listen to the audio question first, then evaluate the responses.

Reference Answer (Gold Standard): {reference_answer}

Response A: {responses_dict['A']}

Response B: {responses_dict['B']}

Response C: {responses_dict['C']}

Instructions:

- Listen to the audio question to understand what was asked.
- Evaluate each response based on correctness and accuracy compared to the reference answer.

- Consider that there may be multiple valid ways to express the same concept.
- Focus on factual accuracy and how well and correctly the response addresses the core question.
- Provide a score from 1–10 for each response (1 = Completely incorrect, 10 = Perfect answer)

Please respond in this exact JSON format:

```
{
  "response_A_score": <score_1_to_10>,
  "response_B_score": <score_1_to_10>,
  "response_C_score": <score_1_to_10>,
  "brief_reasoning": "<brief_explanation_of_your_scoring>"
}
```

B.3 Additional Code and Datasets

The entire code, along with the heuristic-based LALM scripts, is available in the below GitHub repo. All of the datasets used for benchmarking and model training are publicly available on Hugging Face. Access to specific data samples used for validations, such as RAVDESS and DAIC-WOZ samples, can be granted upon request.

GitHub: github.com/DrDongSi/icare/TVR28

Dataset: huggingface.co/TVRRaviteja