Ensemble Listening Models for Voice Intelligence

Introducing Velma

modulate

contact@modulate.ai
modulate.ai

Overview

LLMs write emails, summarize documents, draft code, and answer questions with uncanny fluency. But when used to power AI phone agents or to handle nuanced and potentially risky interactions like customer service calls, account resets, interviews, and more, they quickly run into hard limitations.

LLMs were never designed to understand *how* people speak. Only what we say. They're trained on internet text, not call recordings. They reply to transcripts, not tone.

This fundamental mismatch between LLMs and the multidimensional nature of voice conversations cannot be solved by adding an extra layer to the LLM or "speech-to-speech" agents (which still lose the nuance of tone during their operation). Understanding voice correctly requires a true rethink of how we build AI. LLMs operate on tokens, and voice is too complex to be tokenized.

Getting concrete, LLMs consistently face two major challenges in voice applications:

1. Voice-to-text pipelines throw away the acoustic and behavioral signals that businesses should care about: prosody, hesitation, intensity, the nature of the voice itself (is it synthetic, etc.), and other cues essential for detecting frustration, well-being, and manipulation.
2. LLMs are programmed to return a most confident response, even if it is unverifiable, making them prone to hallucinations.

In creative workflows, that's annoying. In customer support, fraud prevention, safety, or compliance, it's a deal-breaker.

This white paper explores the shortcomings of monolithic LLMs when used to analyze voice and phone calls, proposes a more accurate and more cost-effective alternative in the form of **Ensemble Listening Models (ELM),** and introduces **Modulate's voice-native ELM called Velma.**

Why Voice Intelligence Needs a New Architecture

You're running late to an event and get a quick text from the host - "You still coming?" Are they concerned, just checking in, or annoyed? It's hard to tell, because text is not how human beings convey emotion. We use subtle vocal cues like pitch, energy, pause lengths, and inflection. When we erase all that and limit ourselves to text, as LLMs do, we're losing vital information that makes it not just hard but impossible to understand the speaker's real intent.

By the time an LLM is asked to make a prediction, most of the information needed to make a decision has been discarded – information like emotion, vocal characteristics, urgency, etc. You can't fully fix this problem by sprinkling more transcripts. It's a structural consequence of ignoring the original audio.

The end result is that anything relying on transcripts cannot deliver accurate results when analyzing conversations. But for those using LLMs, this isn't even the only challenge.

Because LLMs are trained to predict likely continuations of text, they will sometimes generate statements that are fluent and plausible but not grounded in the underlying data, also known as hallucinations. Even with careful prompting and guardrails, large models can still invent facts, misattribute intent, or fabricate causal explanations.

Avoiding these issues requires an AI architecture that is deterministic and transparent, and that can supervise generative models like LLMs in real-time. That's not just another LLM, it's something completely new - **an Ensemble Listening Model.**

...the binary grading system [of LLMs] itself still rewards guessing whenever search fails to yield a confident answer.

Why Language Models Hallucinate

Introducing Ensemble Listening Models (ELM)

An ELM is a coordinated ensemble of specialized models organized under a shared orchestration layer. Each model has a clear, narrow mandate, and its outputs are fused in a structured, context-dependent way rather than simply averaged.

The “orchestrator” which balances all these insights also understands context - it can recognize if a background comment is relevant, or notice that a particular emotion fundamentally changes how a recent comment should be understood. ELMs are purpose-built for real-time multi-dimensional intelligence, with an emphasis on transparency, accuracy, cost-effectiveness, and reliability. An Ensemble Listening Model extends the ensemble concept in three important ways:

1. **Embraces heterogeneity**
   Instead of assuming that all base models share the same architecture, an ELM is built around the idea that different tasks (emotion classification, fraud pattern detection, AI-speech identification) are best served by different model families.

2. **Evaluates context over time**
   Each component model evaluates the call at a given point in time - and all the results inform the collective understanding of the type of conversation, which in turn updates how the next part of the conversation is analyzed.

3. **Introduces a multi-tier orchestration layer**
   This key layer does not simply provide averaged predictions, but instead provides reasoning over how different signals interact.

Where traditional approaches to using multiple AI models still only produce a one-dimensional result, an ELM builds a multi-dimensional map of how different behavioral signals rise and fall throughout a conversation, and then synthesizes that map into actionable insights.

ELMs are purpose built for real-time, multi-dimensional intelligence

Velma

Velma is Modulate's enterprise-grade voice intelligence engine, designed to understand what's really happening in voice conversations and not just what words were spoken.

Unlike LLMs that convert speech into text and analyze transcripts with large language models, Velma listens to voice the way humans do. It simultaneously evaluates what is said and how it is said, incorporating tone, emotion, timing, stress, intent, and even signs of deception or synthetic voices. This allows enterprises to move beyond surface-level transcription and gain accurate, actionable insight into real-world conversations.

Velma uses the Ensemble Listening Model architecture, incorporating over 100 specialized models in real time, each focused on a specific signal such as emotion, speaker behavior, escalation risk, fraud indicators, or impersonation. These signals are then combined through an orchestration layer that produces clear, explainable outcomes.

The result is a system purpose-built for enterprise reality:

- Higher accuracy in complex, noisy, emotional conversations
- Transparent, explainable outputs rather than black-box decisions
- Dramatically lower costs than large foundation models at scale

Velma is already deployed across hundreds of millions of conversations for Fortune 500 companies and leading game studios, where it identifies risks like fraud, abuse, dissatisfied customers, policy violations, and malfunctioning AI agents all in real time.

In benchmark testing, Velma outperforms leading foundation models on conversation understanding while being 10–100× more cost-effective, proving that enterprises don't need bigger models, they need better architectures.

In short, Velma turns messy, human voice data into reliable, enterprise-ready intelligence, enabling faster decisions, lower risk, and deeper understanding at a scale traditional AI simply can't handle.

See Velma in Action →