High Sample Rate Evaluation of Audio Playback Quality Using Deep Learning Audition Models

Baicheng Huang[1], Xinyi Pan[2], Haiwei Chai[2] (Oral Presenter), Feng Zhu[2], Dong Liu[2], Xiaoyong Pan[2]

1. *Tsinghua University, Beijing, China*
2. *Conceptual Computing, Cambridge, MA, USA*

Introduction & Motivation

•**The Problem:** Traditional audio evaluation relies on subjective human listening tests, which are:

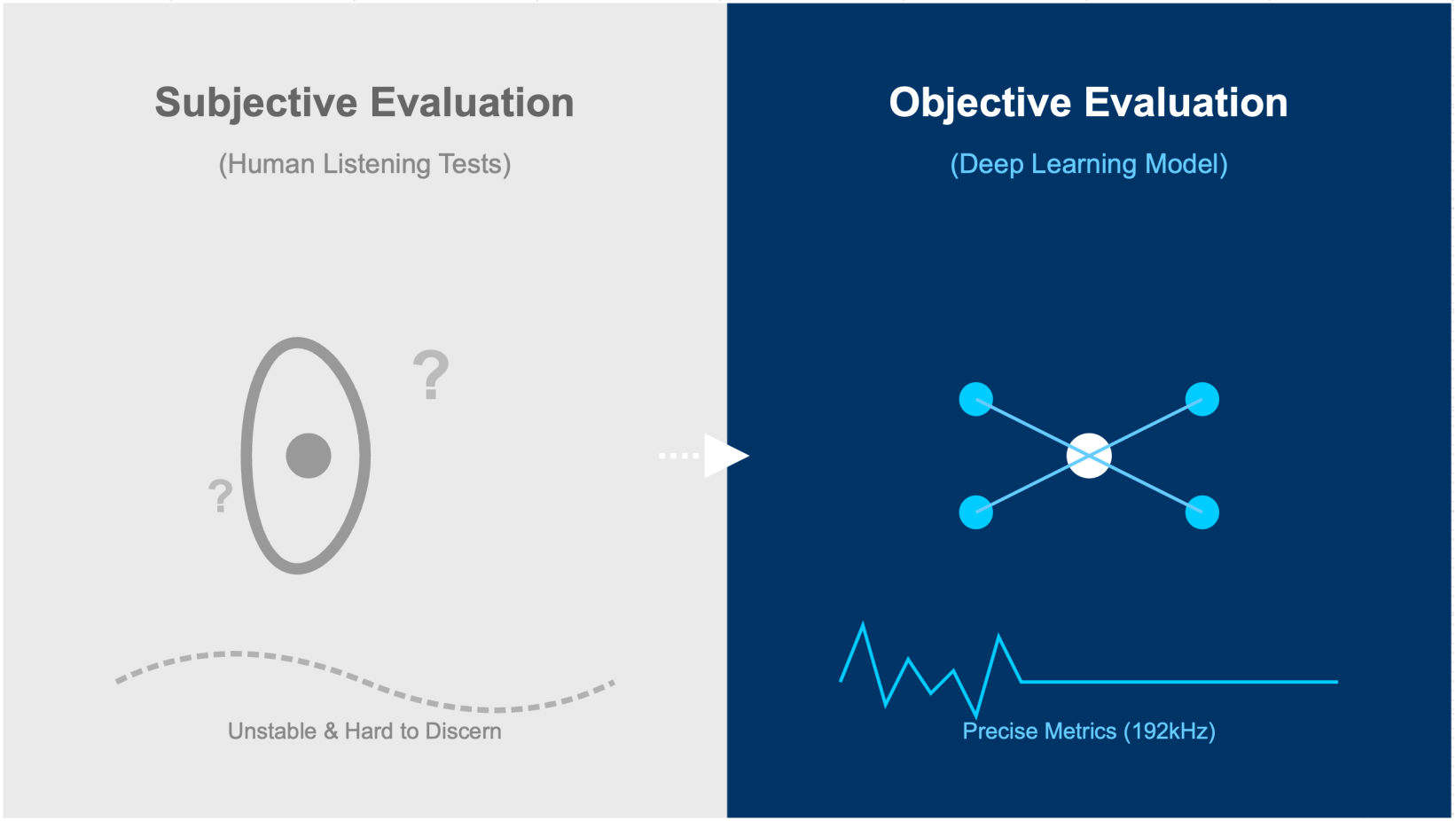
•Inconsistent across different listeners.

•Time-consuming and costly.

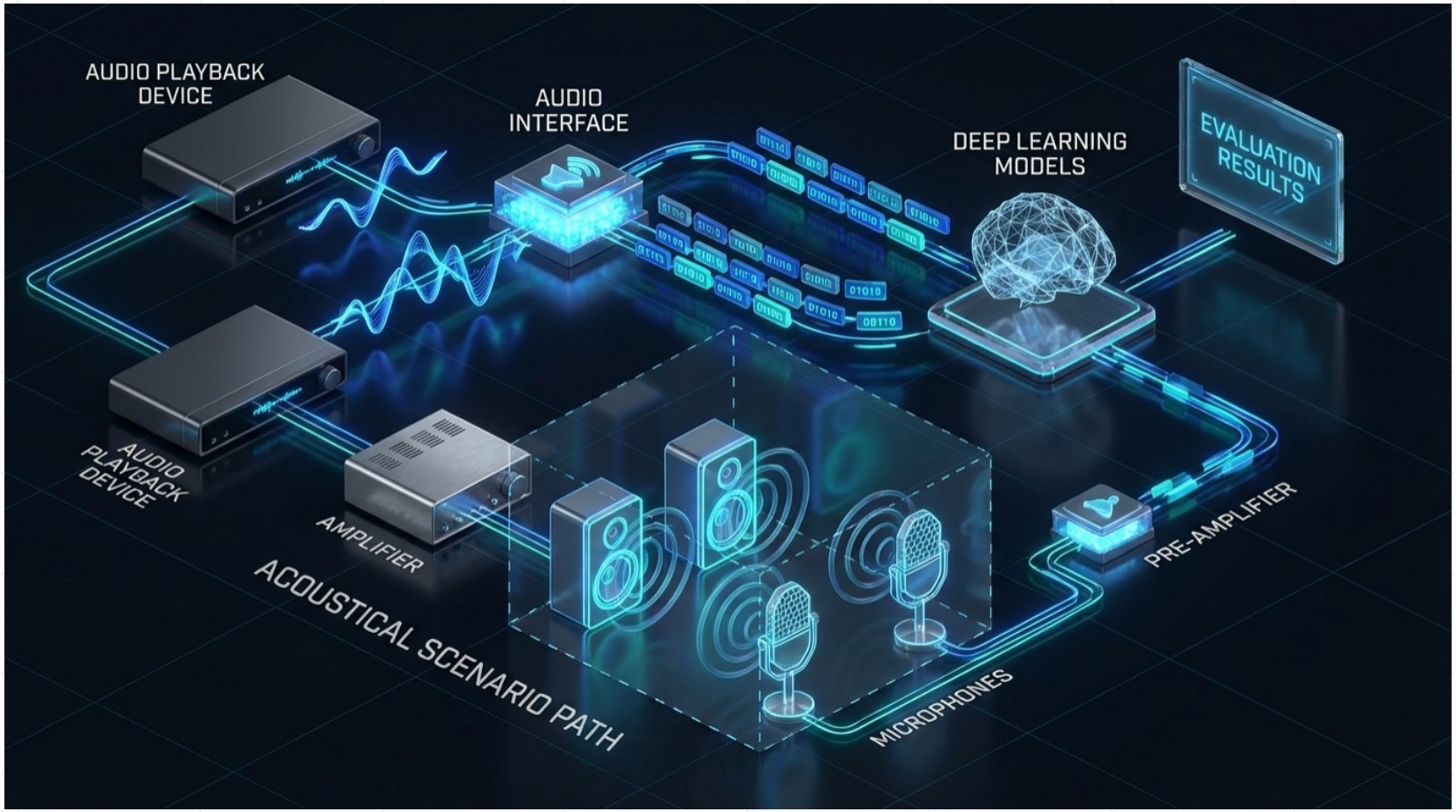
•Limited by human hearing thresholds (subtle differences are hard to discern).

•**The Solution:** Computer Audition Models based on Deep Learning.

•**Research Goal:** Leveraging **High Sample Rate (192.0 kHz, 24 bit)** audio to capture microscopic acoustic details for more accurate automated evaluation.



System Overview



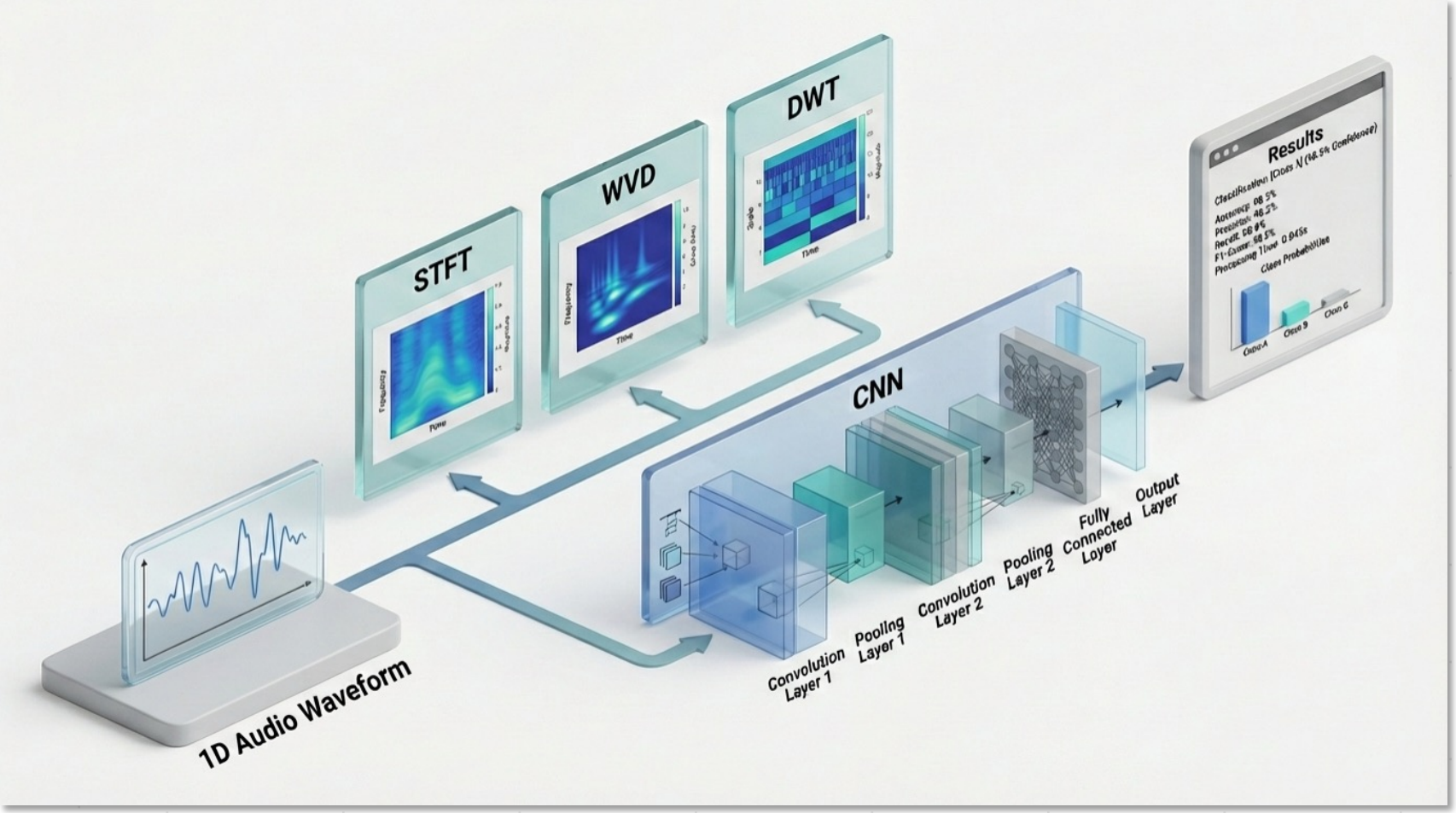
Scenario A: Direct Playback Output

Focuses on the electronic characteristics of the playback device and audio interface.

Scenario B: Acoustic Environment Playback

Includes the influence of amplifiers, speakers, and room acoustics.

Deep Learning Architecture I: CV-Based



•**Concept:** Treating audio as "Auditory Pictures."

STFT

Efficient; intuitive frequency components;

But Fixed resolution; blurs transient details.

WVD

Highest joint resolution; captures self-similarity.

•But Cross-term interference in multi-component signals.

DWT

Multi-resolution; excels at transient capture;

But Basis function (wavelet) selection is critical

Signal Representations (The Inputs)

- **Why 192.0 kHz?** To capture high-frequency harmonics and transient phase information often lost at 44.1 kHz.

- **Feature Engineering:**

- **STFTM:** Magnitude of the Short-Time Fourier Transform.

$$|S_w(u,w)| = \left| \int_{-\infty}^{\infty} s(t)\mathrm{win}(t-u)e^{-jwt}\,dt \right|,$$

- **WVD (Wigner-Ville Distribution):** High-resolution time-frequency representation for transient analysis.

$$W_s(t,w) = \int_{-\infty}^{\infty} s\left(t+\frac{\tau}{2}\right) s^*\left(t-\frac{\tau}{2}\right) \exp\{-iw\tau\}\,d\tau,$$

- **DWT (Discrete Wavelet Transform):** Multi-resolution analysis capturing both coarse and fine acoustic details.

$$T_s(a,b) = w(a)\int_{-\infty}^{\infty} s(t)\psi^*\left(\frac{t-b}{a}\right)dt,$$

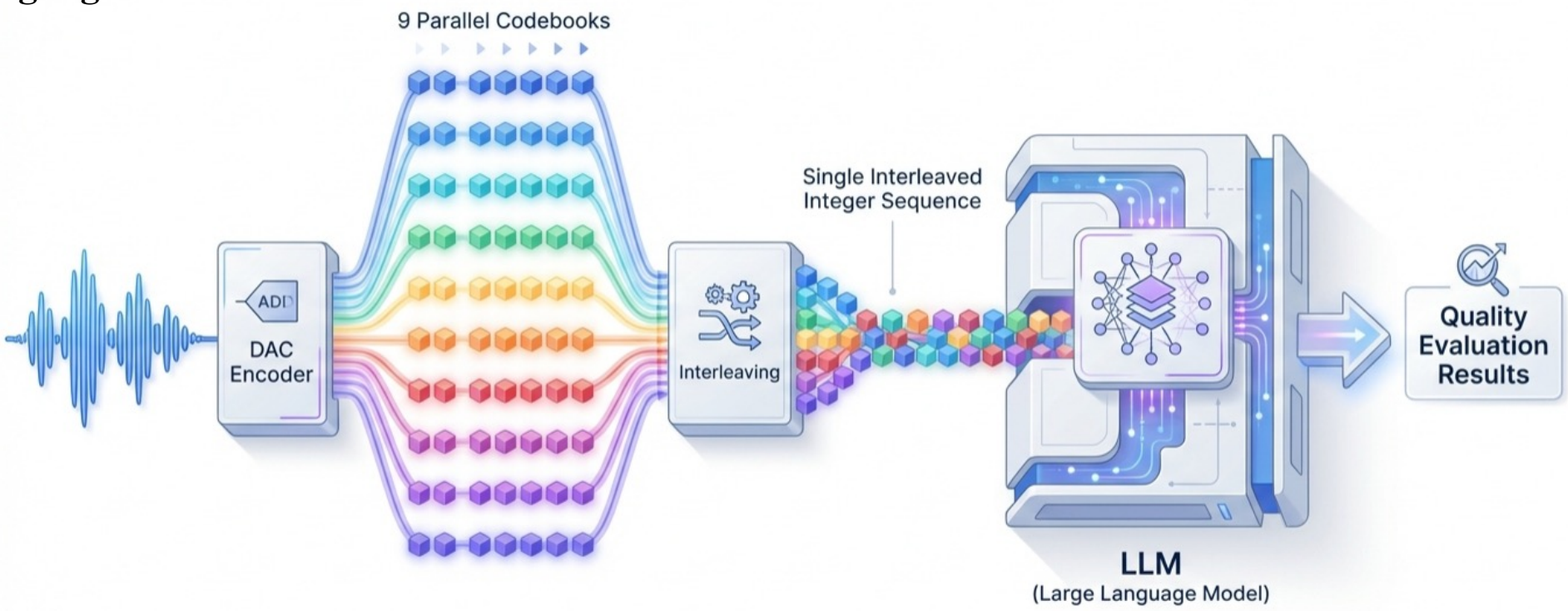
Deep Learning Architecture II: LLM-Based

Core Concept: Audio as a Language

•**Audio Encoding**: Continuous waveforms are discretized using the **Descript Audio Codec (DAC)**.

•**Hierarchical Tokenization**: Utilizes 9 parallel codebooks, generating **9,216 base tokens** in a flat integer space.

•**Interleaving Process**: Parallel codes are flattened into a 1D sequence to simulate "Auditory Language".



Detail: Audio Tokenization & Serialization

Hierarchical Quantization (DAC)

•**Multi-Codebook Structure**: Decomposes each audio frame into **9 parallel discrete codes**.

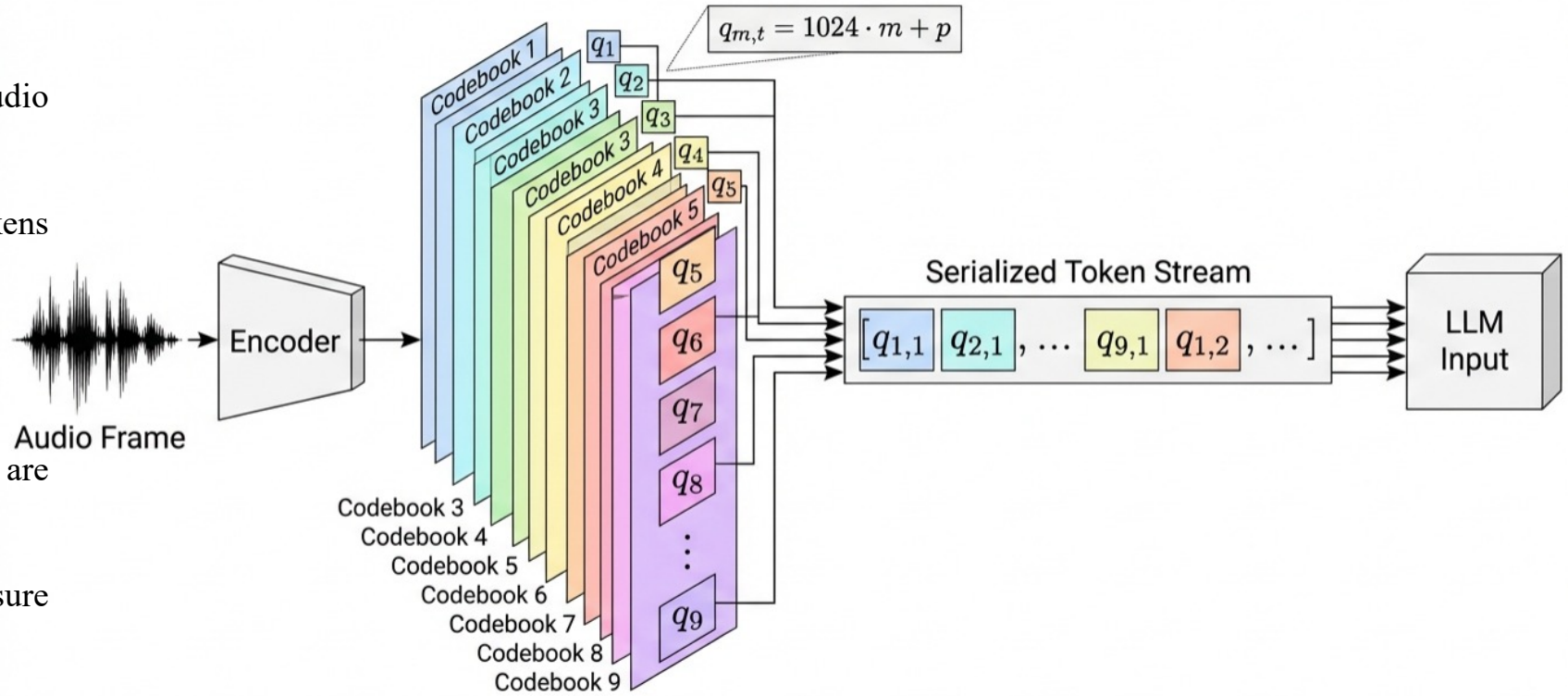
•**Vocabulary Mapping**: Maps 9 codebooks (1,024 tokens each) into a **9,216 non-overlapping integer space**.

1D Sequence Generation

•**Interleaving**: Parallel codes from a single timestep are flattened sequentially.

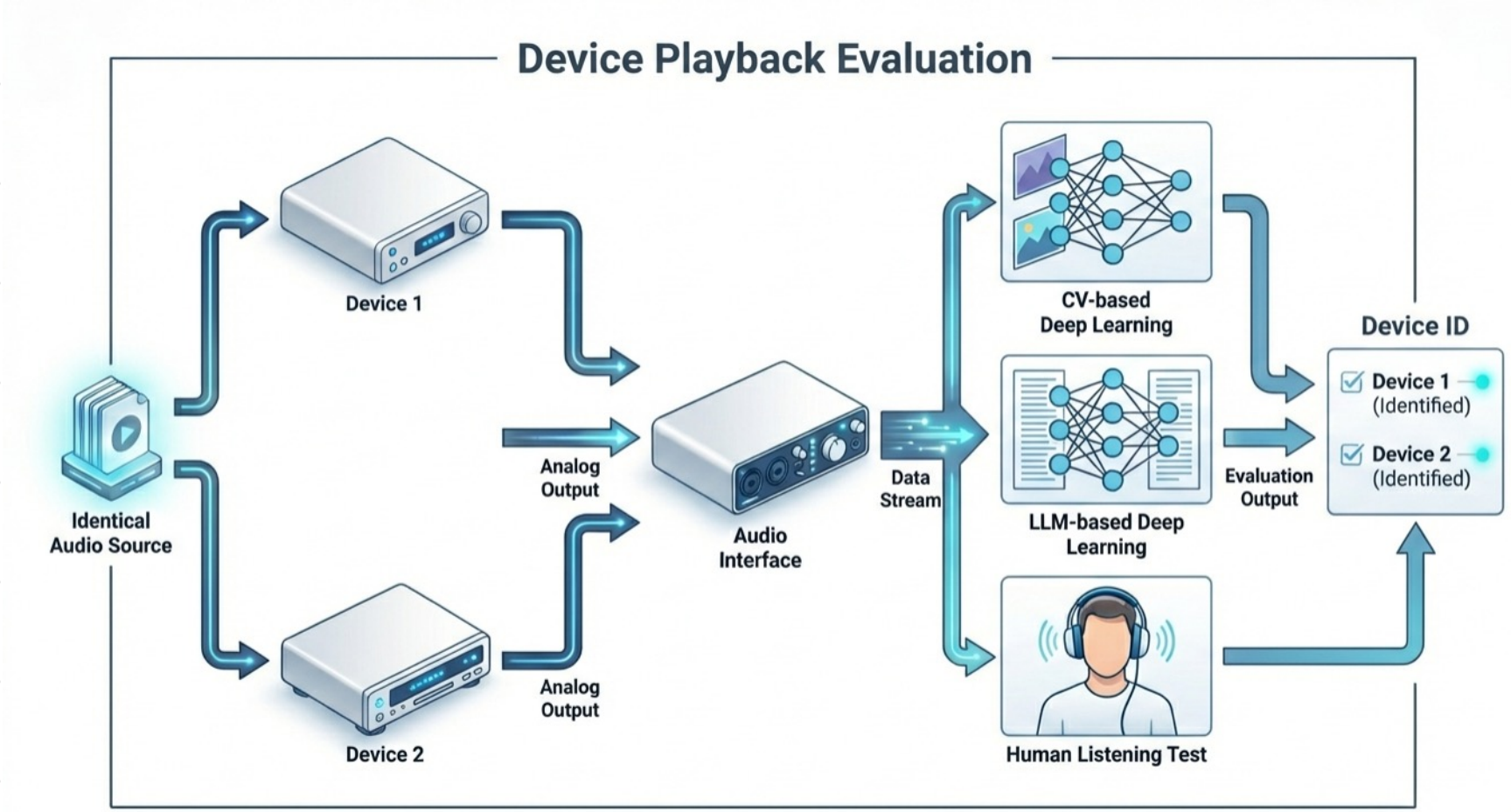
•**Frame Alignment**: Audio is precisely padded to ensure length is a multiple of the model stride (512 samples).

•**Result**: A continuous "token stream" that serves as the final input for LLM training.



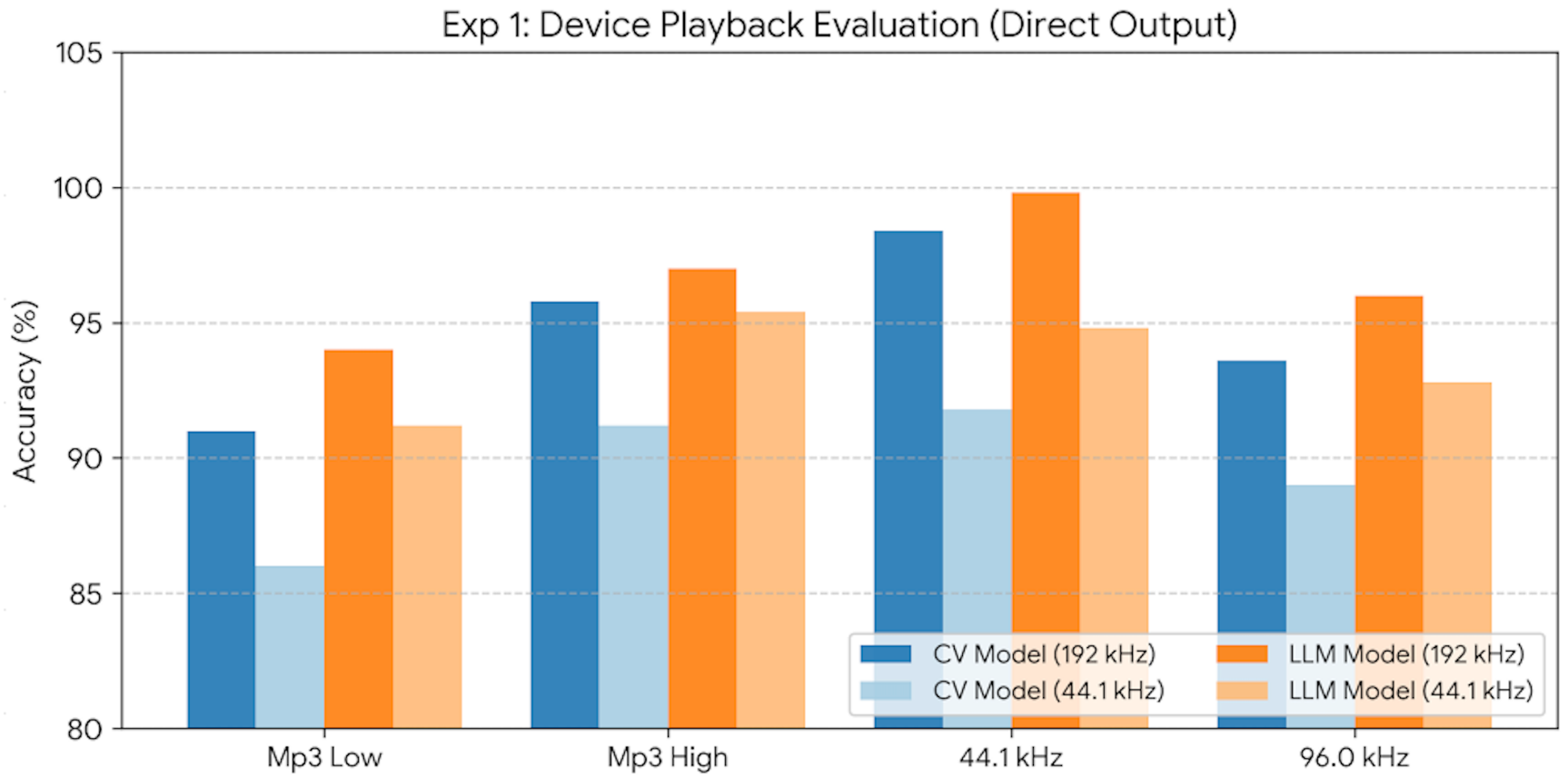
Experimental 1: *Device Playback Evaluation*

This experiment assesses the system's capacity to differentiate between three media players playing identical audio source files.



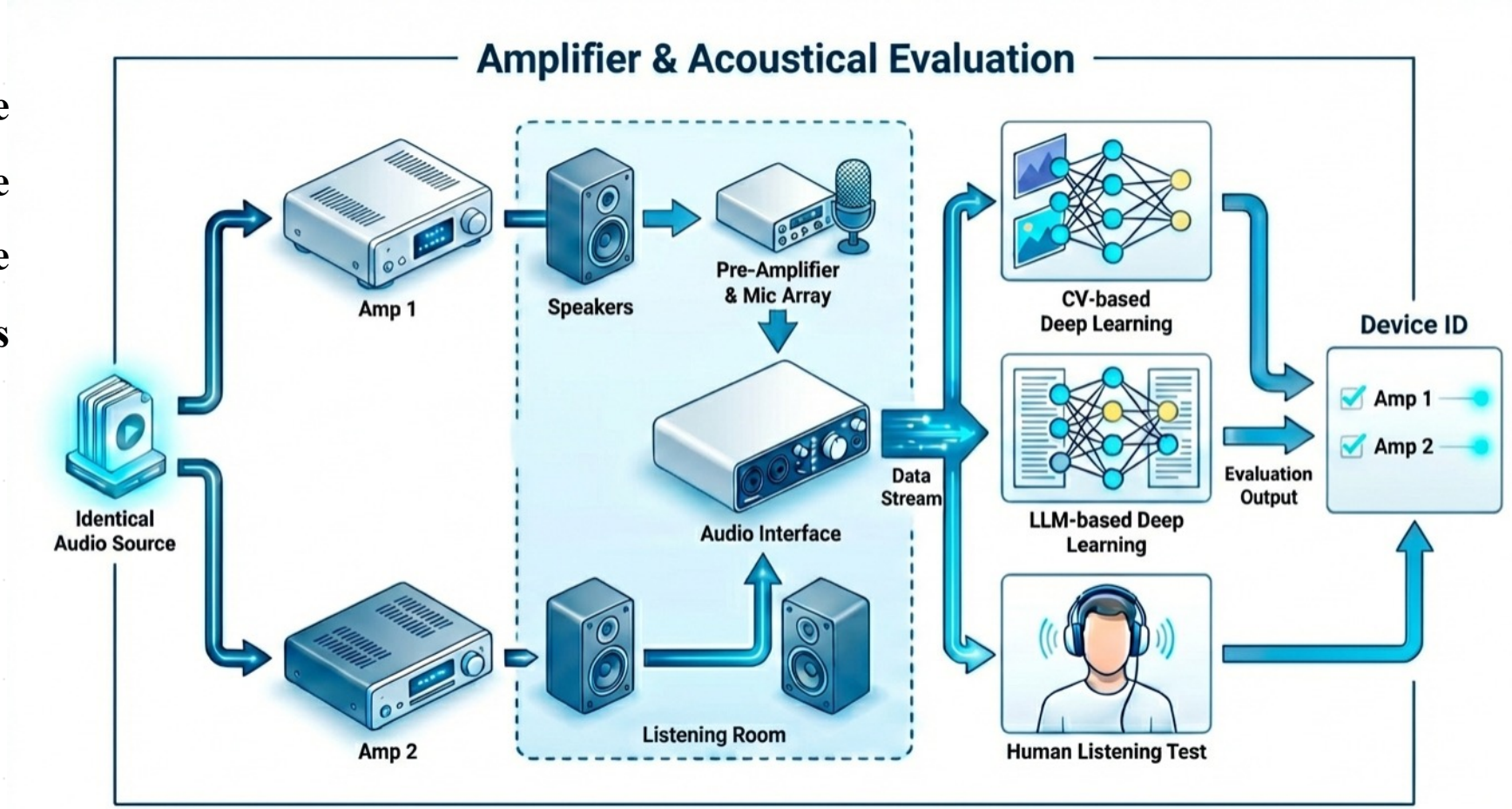
Result Analysis 1: *Device Playback Evaluation*

•**Key Finding:** 192.0 kHz significantly outperforms 44.1 kHz.



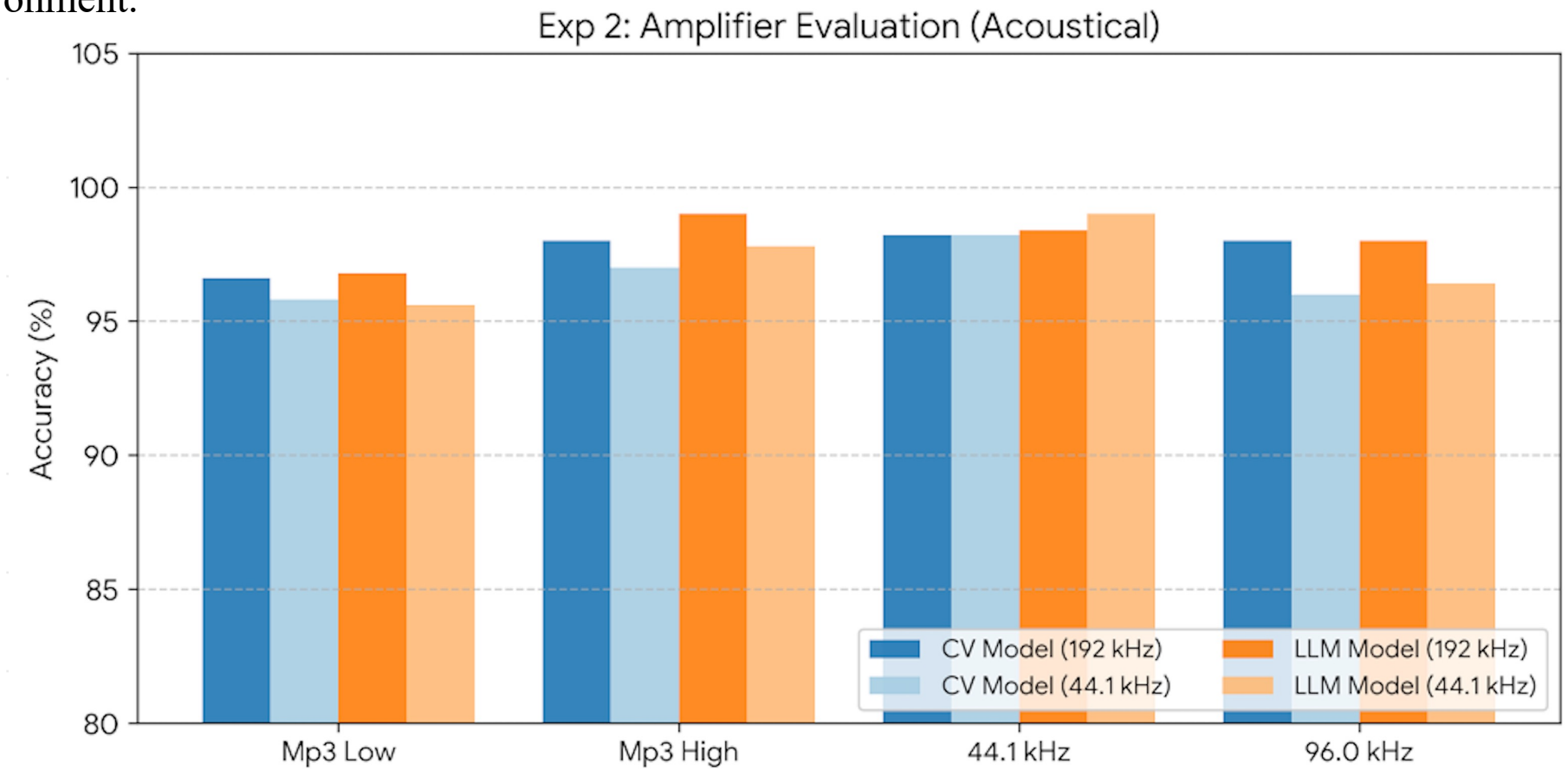
Experimental 2: *Audio Amplifiers*

This experiment compares the acoustical characteristics of three different audio amplifiers when the same signal source and loudspeakers are employed.



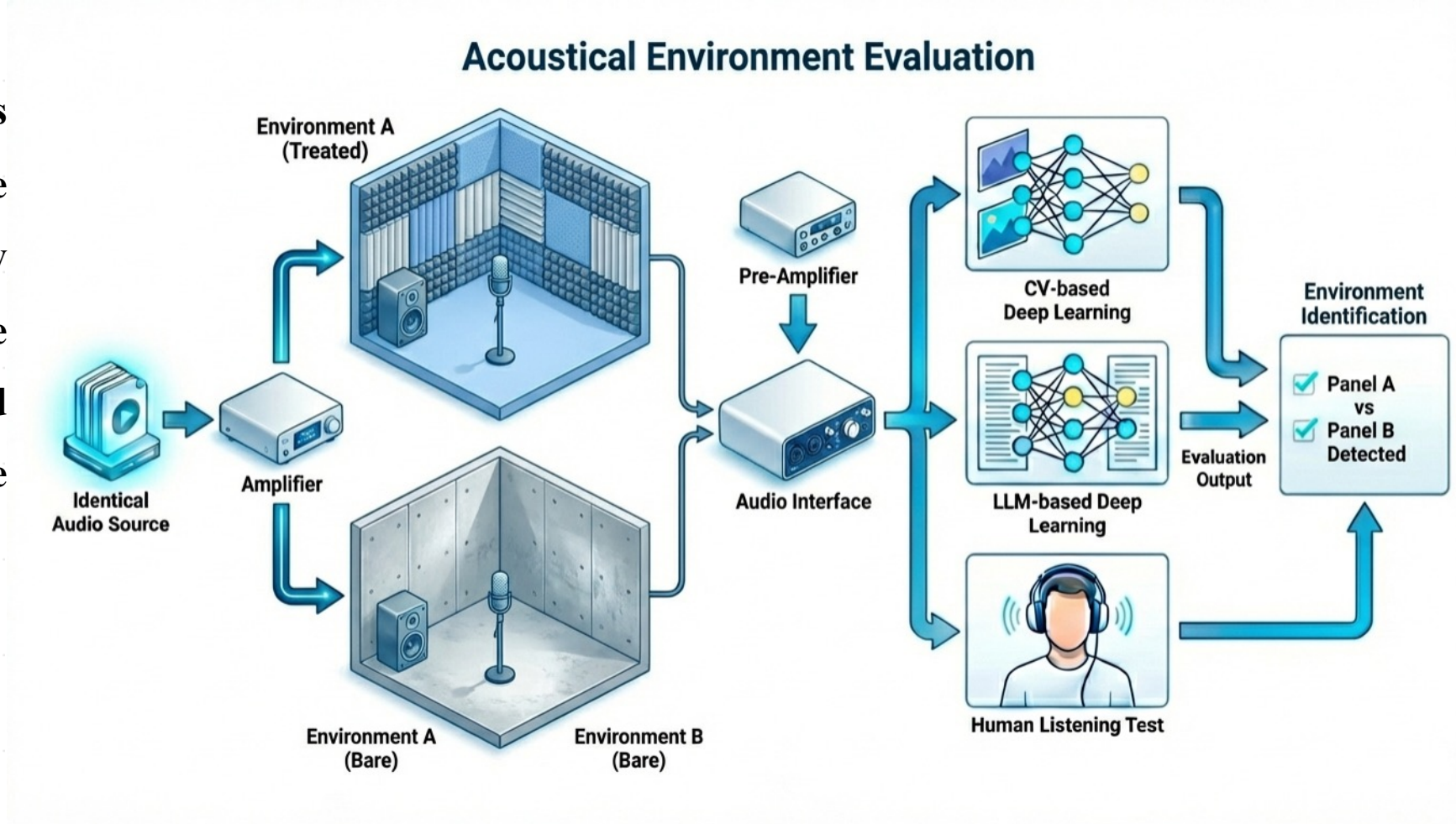
Result Analysis 2: *Audio Amplifiers*

•**Observation:** The system is highly sensitive to acoustic room treatments (e.g., absorption panels).
•**Explanation:** High-frequency phase shifts and reverberation tails at 192.0 kHz provide distinct "fingerprints" for the environment.



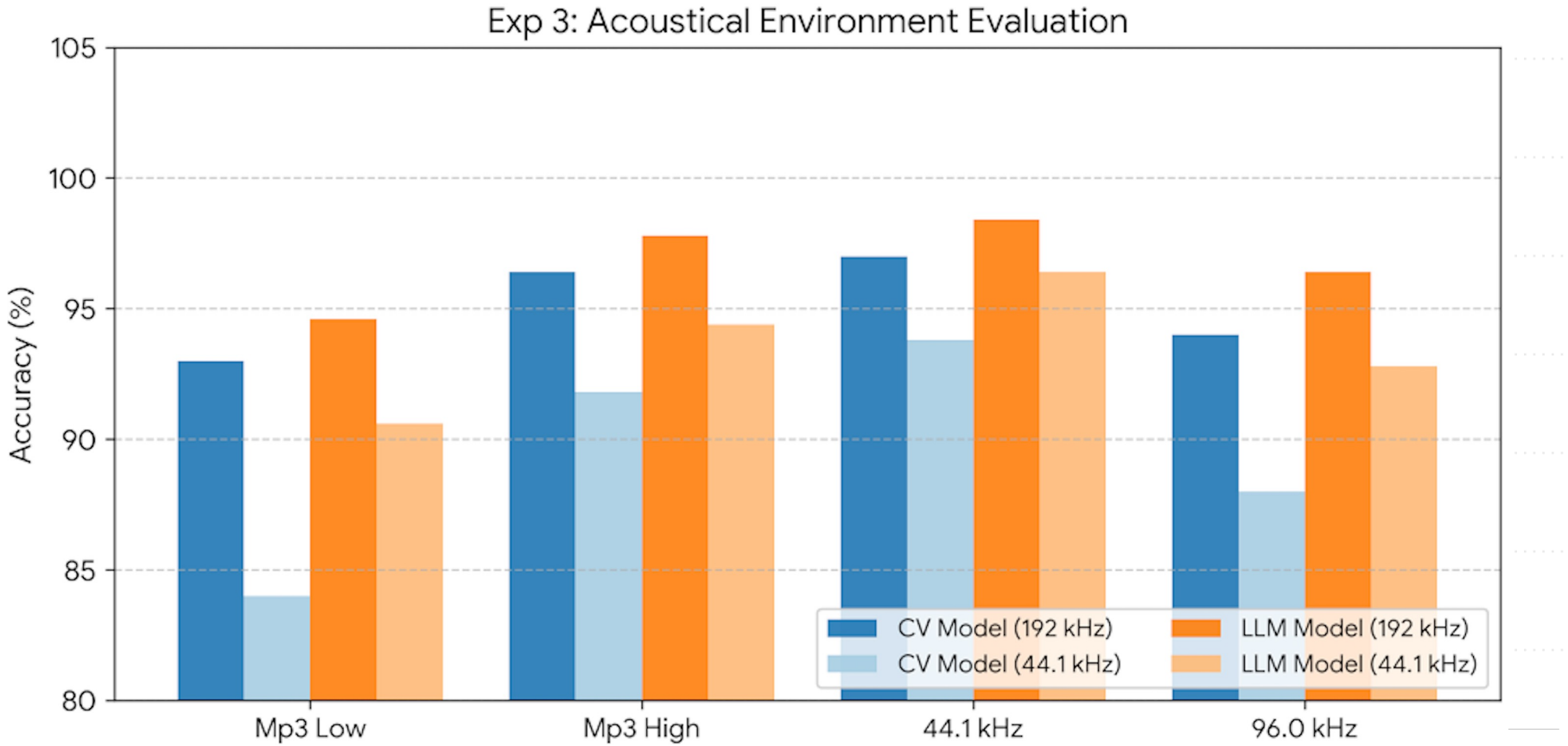
Experimental 3: *Acoustical Environments*

This experiment tests the system's sensitivity to subtle changes in the acoustical environment, specifically by differentiating between the effects of three different sound absorption panels installed in the listening room.



Result Analysis 3: *Acoustical Environments*

•**Comparison:**
•Humans: Struggled with "Difficult" tasks (subtle device differences).
•DL Models: Maintained high stability and accuracy across all difficulty levels.
•**Conclusion:** Deep learning models provide a more objective and robust metric for high-end audio hardware testing.



Conclusion & Future Work

Summary:

Successfully implemented CV and LLM-based audition models.

Proved the necessity of High Sample Rate (192.0 kHz) for automated evaluation.

Future Work: * Expanding the dataset to include more diverse acoustic environments.

Real-time implementation for audio production quality control.

Thank you for your attention!

- **Contact Info: Haiwei Chai (chaihw@conceputing.com)**