UtteranceIQ:A voice-driven, evidence-based interview evaluator with auditable scoring

Kushal Sai Ravindra • University of Colorado Boulder • Kushal.Sair@colorado.edu

University of Colorado Boulder

Motivation

- **Inconsistent evaluation**: two interviewers can rate the same candidate differently without a standardized rubric.
- **Lost technical details**: domain-heavy steps (configs, error handling, metrics) are often missed in notes.
- **Time pressure**: interviewers cannot listen, probe, take clean notes, and score objectively at the same time.
- **Domain depth**: Strong answers contain important technical signals Related to Job domain like system logic, data flow, error handling, and configuration decisions.
- **Auditability**: explainable hiring decisions need evidence, not “gut feeling”.
- **Enterprise constraints**: many orgs avoid LLMs due to privacy/compliance and hallucination risk—rules + database are safer and controllable.
- **Faster, fairer screening:** consistent evidence reduces bias from interviewer mood or style.

References

Link: https://github.com/kusa1379-pixel/UtteranceIQ

Project Goal

- Standardize interviews with a consistent question bank and rubric.
- Create an auto-transcribing answers and capturing structured notes Interview Application.
- Enable evidence-based and auditable evaluation of responses.
- Trigger follow-up questions when key concepts are missing.
- Assess depth using structured answer patterns and measurable impact.

Output: competency scores (1–5), notes, missing signals, and an audit trail

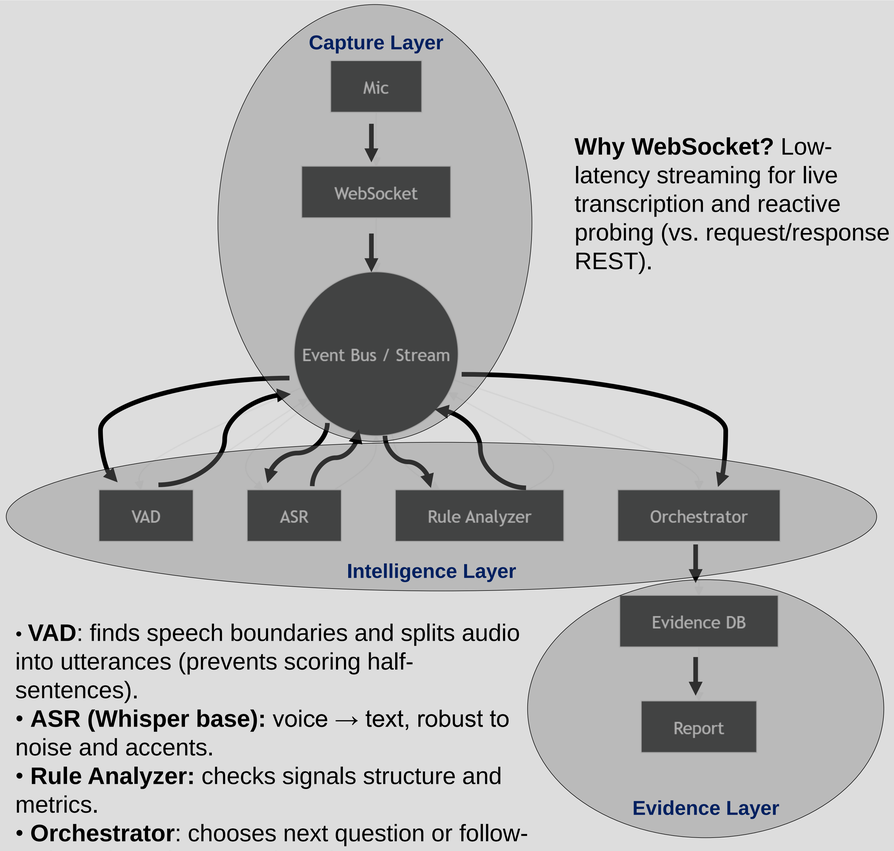
What Makes It Different?

- **Evidence-based scoring**: Scores backed by transcript evidence for clear and reviewable decisions.
- **Domain-aware analysis:** Captures expert reasoning, not just keywords.
- **Controlled & auditable:** Rule/database system ensures predictable and compliant scoring (no LLM dependency).
- **Adaptive probing:** Follow-up questions triggered when answers lack depth, metrics, or required signals.

Database Construction (Signal Library)

- Building a reliable, company-grounded database by capturing **day-to-day annotations** of how employees actually perform each business flow.
- Each flow is broken into **steps**, **actions**, **artifacts**, and **outcomes.**
- **Signal tagging:** Every annotation is tagged to standardized **signals.**
- **Rubric alignment:** Signals are mapped to rubric criteria.

System Overview



Why WebSocket? Low-latency streaming for live transcription and reactive probing (vs. request/response REST).

- **VAD**: finds speech boundaries and splits audio into utterances (prevents scoring half-sentences).
- **ASR (Whisper base):** voice → text, robust to noise and accents.
- **Rule Analyzer:** checks signals structure and metrics.
- **Orchestrator**: chooses next question or follow-up based on triggers and weakest competency.
- **Evidence DB**: stores transcripts, question order, found/missing signals, scores, and follow-ups for auditability.

Sample Evidence Trail

Example record captured for auditing:

Q12 asked → transcript saved → score 4/5
Found: Job Domain Specific Signals
Missing: metrics
Trigger: no_metrics → follow-up asked

Pilot Setup

Item	Pilot / Demo Context
Interview type	Voice-based Oracle HCM mock interviews
Sessions run	15–25 mock sessions
Target roles	ML Engineer / Fast Formula Developer / Reporting Lead / Integrations Specialist
Frontend	FastAPI + WebSocket pipeline
ASR engine	Whisper (base)
DB	SQLite
Typical interview length	20–30 min (≈10 main Q + 2–5 follow-ups

Results

Runtime Performance

Metric	Result (demo/pilot)	Notes
Transcript reaction latency (after pause)	~1.0–2.5 sec	VAD close + ASR time
Follow-up selection time	< 1 sec	Rule-based triggers
VAD end-of-answer detection	0.8–1.2 sec	Silence window; tunable
Transcript usability (quiet room)	~80–90% usable	Enough for signal scoring
Transcript usability (noisy room)	~65–80% usable	Acronyms suffer more
Interview consistency	~80% (deterministic)	Same transcript ⇒ same score

Efficiency vs. Real Interview

Metric	Traditional interview	UtteranceIQ (voice + rules)
Interviewer typing during interview	20–35% of time	5–10% (monitoring)
Time to produce debrief	10–25 min	30–90 sec + 2–5 min review
Missed details in notes	Medium–High	Low–Medium
Consistency across interviewers	Medium	High

Impact & Future Work

Turning a person into a score can reduce candidates to “signals,” making the process feel less respectful.

- **Style over substance**: RST coherence, rhetorical relations.
- Improve ASR for Signal acronyms using a domain lexicon and fuzzy matching.
- Expand signals beyond keywords (phrases, clusters, error-signature patterns) while staying no-LLM.
- Add a hands-on practical round.
- Extend to video screening alongside audio.