REQUEST FOR PROPOSAL (RFP)

Speaker Attribution Data Annotation of ORF Voice Recordings

Closing Date for Submission

15th September, 2026, at 23:59 PM

I. INTRODUCTION

LEHS is a charitable organization in India whose purpose is to offer basic health and education for the poor. LEHS in furtherance of charitable objectives through its flagship programs, Wadhwani AI, which aims to build equitable and sustainable systems by making quality primary healthcare available and accessible to the underserved population, and to bring the benefits of modern AI technology to underserved populations by building and deploying AI solutions for social impact across domains such as healthcare, agriculture, governance, and education in India. LEHS aims to promote the integration of technologies, particularly in emerging domains like artificial intelligence and innovations into the Indian mainstream primary healthcare, education, and agriculture systems through a partnership with the State and National Government, apex institutions, international agencies, and private sector partners e.g. innovators, social enterprises and other ecosystem contributors in line with its stated objectives for the betterment of society particularly focusing on projects of national and social significance. LEHS also undertakes rigorous monitoring, evaluation, and learning (MEL) activities to assess the impact, usability, and scalability of different programmatic interventions.

Wadhwani AI, a unit of LEHS, focuses on developing, deploying, and evaluating artificial intelligence solutions to address critical social challenges in India, particularly in domains such as healthcare, agriculture, and education.

II. BACKGROUND

A significant number of students in Indian public schools (Grades 1–8) lack foundational reading skills, impacting their academic progression across all subjects. This challenge is compounded by the absence of scalable, data-driven tools for assessing reading in diverse languages and for identifying specific proficiency levels to enable targeted interventions.

In response, LEHS (Wadhwani AI) has developed an AI-powered tool that **automates the assessment of foundational reading skills**. The tool leverages highly contextualized and fine-tuned Automated Speech Recognition (ASR) models to analyze word and paragraph reading, generating key performance metrics and identifying specific reading miscues. Our solution also applies an intelligent layer to cohort students into reading levels, providing teachers with actionable insights for planning targeted classroom interventions.

To date, LEHS (Wadhwani AI) has assessed **over 8 million students in Gujarat and Rajasthan** for the Gujarati and Hindi languages**.** When the assessments are done in a

classroom environment, there are many issues with the quality of the audios that teachers use for child's assessment as ASR catches multiple speakers including adult reading in place of a child. We want to improve our mechanism to detect such speakers and train models that can perform well when people are speaking over each other, especially adults.

III. PROJECT OBJECTIVES

This RFP is issued to solicit bids from experienced data annotation agencies to generate high-reliability ground truth data for speaker attribution within Oral Reading Fluency (ORF) assessment audios. This dataset will support the development of models capable of detecting multiple speakers, specifically substitution (an adult reading in place of the child), prompting (an adult providing assistance to the child) and overlap (an adult speaking over child reading).

IV. SCOPE OF WORK

Summary of Activities

S.No	Activity Description	Timeline
1	Co-Working & Onboarding Phase	**2nd week of October 2026**
2	Data Annotation for ASR: Create speaker attribution data following Wadhwani AI's data annotation protocol.	**October, November 2026**

- **Co-Working & Onboarding Phase**

  The selected partner shall participate in a structured co-working and onboarding phase during the first couple of days following contract finalization. Participation in this phase is mandatory and is a prerequisite for commencing full-scale annotation.

  The objective of this phase is to establish end-to-end operational alignment between Wadhwani AI and the partner prior to full-scale annotation, **including a demonstration of the annotation tool by the partner**, an **overview of the annotation protocol**, and **alignment on quality assurance and review processes** to ensure high quality annotation output from the partner. The tool demonstration must confirm that the annotators can annotate two or more speech segments over the same time range, since overlapping speakers are tagged as separate segments.

Full-scale annotation may commence only after formal readiness sign-off by Wadhwani AI. Failure to satisfactorily complete this phase may result in delays to full scale annotation or reconsideration of the engagement.

- **Data Annotation for Speaker Attribution**

Definition of "Adult Speech":

Every speech segment of human speech in the chunk needs to be annotated, whether or not the words are intelligible. The single most important judgment for this exercise is whether a speech segment was spoken by an adult.

Rule: mark a speech segment as adult only when there is no doubt that the speaker is an adult. This bar is about the identity of the speaker, not about the words. A speech segment can be unintelligible, faint, distant, or whispered and still be marked adult, as long as it is unambiguously an adult who is speaking. Loudness, distance, and clarity are recorded separately in the speech prominence field and must not influence the adult decision.

Tag as Adult (examples):

- **Direct instruction:** the adult reads words, sentences, or instructions to the child.
- **Substitution:** the adult reads the assessment text as if they were the test-taker.
- **Prompting or repeating:** the adult provides the correct word or phrase, or the child repeats after the adult.
- **Background or incidental adult speech,** including cross-talk and shouting, as long as it is unambiguously an adult. Where it sits acoustically is recorded in Speech Prominence, not here.

Do not tag as Adult:

- Any speech segment where you cannot be sure an adult is speaking. If it is borderline between an adult and an older child, do not force the call. Label it Uncertain or Child, but sounds like an adult in the Speaker Label field (see the annotation fields below).
- Very short or unclear sounds that cannot be confidently attributed to an adult, even if they sound adult-like. If it is human speech, it still gets a timestamp and is labeled Uncertain unless you are sure that the utterance was from an adult.

3. Annotation Workflow and Fields

Annotation proceeds in two passes over each 20 second chunk.

- **Segment the speech:** A speech segment is a continuous stretch of speech from a single speaker, carrying one speaker label and one interaction type. Start a new speech segment whenever the speaker changes or the interaction type changes. When a single speaker's interaction type changes mid-utterance, for example correcting, then prompting, then praising in one continuous sentence, split at the nearest natural pause or word boundary. When two or more people speak at the same time, tag each voice as its own speech segment over the respective time range. **Silence and pure non-speech noise are left untagged.**

- **Label each segment.** For every speech segment, first set the Speaker Label. Make sure you are certain when you tag the speaker label as Adult. Then set the Interaction Type, Speech Prominence and Whispering. Speech prominence and whispering are recorded for each speech segment but do not, on their own, start a new speech segment, unless prominence changes clearly and stays changed.

Rule: Interaction type and speech prominence and whispering are tagged for every speech segment regardless of the speaker label.

Annotation Field	Description	Response Options
Speech Segments	Precise start and end timestamps of the speech segment.	Timestamps in seconds (e.g., 04.20 to 06.50). Do not tag silence or pure non-speech noise.
Speaker Label	Who is speaking: Main label for speaker identity. **Marked for every speech segment.**	1. **Adult**: clearly identifiable as an adult speaker. 2. **Child**: clearly identifiable as a child speaker. 3. **Child, but sounds like an adult**: identifiable from context as the child, but with a deeper, adult-like voice. 4. **Adult, but sounds like a child**: identifiable from context as an adult, but with a higher, child-like voice. 5. **Uncertain**: cannot tell, even from context.
Interaction Type	What kind of speech is this? **Marked for every speech segment.**	**1. Assessment reading**: reading the actual assessment text. **2. Prompting**: taking turns with the reader, word by word or phrase by phrase. **3. Instruction & Feedback**: telling the reader or other children in the classroom what to do,

Annotation Field	Description	Response Options
		encouraging, praising or correcting pronunciation, and not reading the assessment text. **4. Noise or cross-talk**: any other speech, such as background talk or shouting.
Speech Prominence	How the speech sounds, depending on where the speaker is with respect to the mic. **Marked for every speech segment.**	**1. Foreground**: clearly audible speech in the foreground, close to the mic. **2. Background:** clearly audible speech in the background but away from the mic. **3. Faint**: barely audible or muffled, but still clearly human speech.
Whispering	Whether the speech is whispered or not **Marked for every speech segment.**	**Yes:** The speaker whispers while speaking. **No:** The speaker does not whisper and speaks normally.

4. Handling Ambiguous Scenarios

These rules standardize the hard cases and keep the adult class clean.

- **Voices that sound like another age-group:**
  a. Child sounding like an adult: If the primary respondent is an older child or has a deep voice, use "Child, but sounds like an adult". This can happen with older boys and sometimes even girls when they are 12-15 yrs of age, tend to sound like 18-21 yr olds. In this case, your human ear judges that this is a child based on the context of the assessment but the voice sounds like an adult's voice.
  b. Adult trying to sound like a child: If you feel like an adult is reading in place of the student, and trying to imitate the sound and style of a child's voice and how a child might read, use "Adult, but sounds like a child". In this case, your human ear judges that this is an adult trying to sound like a child. If you still cannot tell whether the reader is an adult or a child, use "Uncertain" rather than guessing.
- **Be certain before marking an adult.** Speaker identity is the most important judgment in this protocol. Mark a speech segment as Adult only when you are sure the speaker is an adult.

Imp Note:

- Annotators work on **fixed-length 20 second chunks** (provided to them) and mark every boundary for the speech segments from scratch.
- **Annotate every stretch of human speech**, whether or not the words are intelligible. Leave silence and pure non-speech noise untagged.
- The single most important judgment is whether a segment was spoken by an adult. **Mark Adult only when you are sure.** When in doubt, do not force the call.
- The aim is to record **who is speaking**, adult or child, and **what kind of speech it is**, so that adult assessment reading and word-by-word prompting can be told apart from ordinary instruction, feedback, or a stray background voice.
- A speech segment is one speaker with one interaction type. Start a new segment whenever the speaker or the interaction type changes.
- When two or more people speak at the same time, tag each voice as its own segment over the same time range.
- Adult and non-primary speech ranges from close reading to distant prompting to background cross-talk. Pay close attention to these variations and record them in speech prominence and whispering. They describe how the speech sounds and do not, on their own, start a new segment.

Annotation Logistics

This annotation process should follow best practices for conducting speech annotations, our recommendations include:

- Annotators should use headphones or earphones (not loudspeakers).
- Annotators should complete the task in a reasonably quiet environment.
- Playback volume should be set to a comfortable listening level and kept consistent.
- Annotators should complete a short training and qualification step in annotating audios having different speakers in the classroom environment.

Other Important Considerations

Audio Duration

We will share **20-second audio** speech chunks.

Total Data Size

We will be aiming for **100 hours of annotated data**. With 20-second clips, this corresponds to around **18000 audio clips**.

Monitoring Annotation Quality

We will begin with a pilot batch of approximately 5 hours of audio. The selected partner will be expected to provide the raw annotations for this batch along with a pilot report.

The pilot report should cover:

- **Process adherence** — how each element of the proposed process described in the Technical Proposal (Section 1.2) was implemented in practice, including annotator training, equipment compliance, guideline adherence, and independent scoring.
- **Quality monitoring** — how annotation quality was tracked during the pilot, how annotator performance was evaluated, and any corrective actions taken.
- **Challenges and resolutions** — any significant challenges encountered during the pilot and how they were addressed.

Once annotations and the report are received, we will manually review a random subset to assess annotation quality.

If annotation quality is not up to the mark at this stage, we may ask the selected partner to revise their annotation protocols based on specific feedback before proceeding to the full dataset. The selected partner is expected to **submit the first batch of 5 hours within the first week of finalisation of the contract.**

Payment eligibility will be linked to the successful submission of the entire scope of work. The partner may raise a single consolidated invoice after delivery. Additionally, the partner will be required to participate in weekly progress reviews to ensure timely completion and adherence to quality standards.

V. TIMELINE AND ROUNDS

S.No	Activity	Date
1	Release of the Request for Proposal (RFP)	Sept 2nd, 2026
2	Q&A and Clarifications	Sept 8th, 2026
3	Last date for submission of the complete proposal	Sept 15th, 2026

VI. REQUIRED DELIVERABLES

The selected partner will be responsible for delivering the following within the project timelines of within **2 Months** and quality standards:

- **Pilot report and pilot annotations: The report and the annotations for the pilot batch of 5 hours.**
- **JSON file for the annotations: The format of the annotation is provided in the Annotation Format section below.**
- **A Quality Justification report on annotation accuracy and process followed after completing all the annotations.**

VII. PROPOSAL SUBMISSION GUIDELINES

Your proposal must include:

1. **Technical Proposal**
    - **Understanding of Scope, Deliverables, and Problem Statement:** Demonstrate understanding of the problem we are solving with these annotations, the full scope of work, and the expected deliverables.
    - **Process and Tools**
        1. Annotator selection criteria
        2. Equipment and listening environment standards
        3. Annotator training process: how annotators are trained to understand each type of annotation and the differences between them
        4. Annotation tooling used
        5. How annotator fatigue is managed
        6. How annotation quality is defined and what mechanisms are used to monitor annotation quality continuously during the project.
        7. How underperforming annotators are identified and addressed
    - **Team Composition and Qualifications:** Details of the team assigned to this project, their relevant experience, and role allocation.
    - **Quality Assurance Mechanisms:**
        1. Remediation process: what happens when annotation accuracy falls below acceptable levels, including the process for how affected data is re-annotated.
        2. Reporting cadence: how frequently the partner will communicate with us during the engagement (daily, weekly, or milestone-based).
        3. Reporting content: what specific metrics or summaries will be shared with us on a regular basis (e.g., annotation throughput, flagged files, annotator performance summaries) and in the pilot and final reports.
    - **Data Policy:** partners must describe their data handling policy with respect to our data, including storage location, access controls, retention period, deletion

process upon project completion, and any subcontracting or third-party access to the data.

2. **Financial Proposal**
    - Detailed cost breakdown
    - Schedule of Payments Linked to Milestones Achieved
    - Applicable taxes and contingencies

3. **Organizational Profile**
    - Relevant project experience
    - Sample of similar work

4. **Signed NDA:** Partners are expected to download and update the Organisation Name, registration no., Address in the NDA here, provide Authorised Signatory and attach signed NDA with your proposal. **Without Signed NDA, Proposal will be invalidated.**

5. **Annotation of Sample Audios:** Partners are provided with a sample set of audio files via Google Drive. Along with their proposals, they must return completed annotations for this sample set as part of their proposal, following the definitions provided above. The format for the annotations is provided in the Annotation Format section. **Without Sample annotations in required format, Proposal will be invalidated.**

Sample Evaluation Criteria

Partners are required to submit completed annotations for the sample audio set provided. These annotations will be evaluated on annotation accuracy annotations calculated against golden ground truth prepared internally. This measures how closely the partner's annotations align with the expected annotators, and is an indicator of how well annotators have understood and applied the metric definitions and score descriptors.

Criteria for Application

Applicants must meet the following criteria.

- **Mandatory:**
    - Demonstrated team capacity to annotate audio clips based on metrics
    - Clear and demonstrable quality check process
    - Submission of evaluation samples (Proposal without sample annotations will be rejected)
- **Preferred:** Prior experience in annotating multilingual audio data, including languages such as Gujarati, Hindi, English, Odia, Telugu, etc.

Submission Details

- All proposals to this RFP must be received no later than **<Sept 15th>.** The proposal should be submitted only through e-mail in PDF format addressed to The Procurement Team at the below-given e-mail id: rfp.lehs@wadhwaniai.org. The email's subject line must contain the reference number and title of the RFP: **Speaker Attribution Data Annotation of ORF Voice Recordings**

- Any proposals received by Wadhwani-AI after the deadline for submission of proposals prescribed in the timeline of this document are liable to be rejected.

VIII. EVALUATION CRITERIA FOR TECHNICAL PROPOSAL

Criteria	**Weightage**
Understanding the scope	5%
Technical soundness of approach and samples received (tools or annotated audio)	35%
Data Quality Assurance	10%
Work plan and staffing	15%
Credentials of the organization in annotating speaker attribution in messy audio data	25%
Credentials of the proposed team	10%
Technical Total	**100%**

Only participants in the RFP qualifying on the technical evaluation will be eligible for the financial evaluation. Financial proposals will be evaluated on criteria including total cost, cost breakdown, cost reasonableness, cost-effectiveness, payment terms, and financial stability of the participating organization. Also, the participants should show a demonstrated compliance with the provisions of the Digital Personal Data Protection (DPDP) Act, 2023, ensuring responsible collection, storage, and use of the students' personal data.

IX. MANDATORY ENCLOSURE

Please submit the following as enclosures or attachments with your quotation. Bidders must provide all the information requested above in section VIII. Quotations that do not provide the required information and certificates, or do not follow the submission requirements, may not be reviewed.

- Company Profile, testimonials, work completion reports, purchase orders, and reference (NGOs).
- NGOs need to provide a valid CSR-1 issued in favor of the NGO by the MCA.
- GST registration certificate.
- PAN & TAN registration.
- Audited Financials for the last three years.
- Income Tax Return for the last three years.
- Cancelled cheque.
- Copy of registration documents/certificate and the most recent renewal as a legal entity.
- Preferably a Private Ltd company and has sound business records.
- Authorized signatory letter.
- Conflict of interest declaration.
- Contact person details.
- Memorandum of Association & Articles of Association (or Trust Deed / Bye-laws as applicable).
- List of Governing Body Members/Board Members of the organization.
- 12A and 80G Registration Certificates and Revalidation Document.
- PF, ESI, and Professional Tax Registration Certificates (if applicable).
- Finance Policy, HR Policy, Anti-Sexual Harassment Policy, Procurement Policy, and any other relevant internal policies.
- List of Pending Litigation/Cases against the Entity (if any).
- Latest TDS Return filed (Form 26Q and 24Q).
- GSTR filings for the last 3 returns (if applicable).
- Sample Audited Utilization Certificate (UC) submitted to any donor for CSR Funding.
- Audit Report in Form 10B/BB for the last 3 years.
- Foreign Contribution (FC) Audited Financials for the last 3 years (if applicable).
- FCRA Returns (FC-4) for the last 3 years (if applicable).

Note: Only shortlisted partners will be contacted for presentations or negotiations. If you do not hear from us within two weeks of submission, consider your proposal not selected. LEHS reserves the right to reject any or all proposals and to negotiate terms and conditions with the chosen partner.

X. KEY TERMS AND DEFINITIONS

Term	Definition
Speech Segment	A speech segment is a continuous stretch of speech from a single speaker, carrying one speaker label and one interaction type.
Adult	A speaker who is unambiguously identifiable as an adult.
Child	The student who is clearly identifiable as a child speaker.
Child but sounds like an adult	Identifiable from context as the child, but with a deeper, adult-like voice.
Adult but sounds like a child	Identifiable from context as an adult, but with a higher, child-like voice.
Interaction Type	A classification of the speech based on its nature, such as assessment reading, prompting or instruction.
Speech Prominence	How the speech sounds based on loudness and distance from the mic.

XI. ANNOTATION FORMAT

```
{
  "audio_file_name_1": [
    {
      "start_timestamp": 4.20, # Speech Segment 1
      "end_timestamp": 6.40,
      "speaker_label": "ADULT / CHILD / CHILD_SOUNDS_ADULT /
ADULT_SOUNDS_CHILD / UNCERTAIN",
      "interaction_type": "ASSESSMENT / PROMPTING / INSTRUCTION_FEEDBACK /
NOISE_CROSSTALK",
      "speech_prominence": "FOREGROUND / BACKGROUND / FAINT",
        "whispering": "YES / NO"
    },
```

```
    .
    .
    .
    {
      "start_timestamp": 5.74, # Speech Segment N
      "end_timestamp": 9.21,
      "speaker_label": "ADULT / CHILD / CHILD_SOUNDS_ADULT /
ADULT_SOUNDS_CHILD / UNCERTAIN",
      "interaction_type": "ASSESSMENT / PROMPTING / INSTRUCTION_FEEDBACK /
NOISE_CROSSTALK",
      "speech_prominence": "FOREGROUND / BACKGROUND / FAINT",
        "whispering": "YES / NO"
    }
  ],
  .
  .
  .
  "audio_file_name_n": [
    {
      "start_timestamp": 4.20, # Speech Segment 1
      "end_timestamp": 6.40,
      "speaker_label": "ADULT / CHILD / CHILD_SOUNDS_ADULT /
ADULT_SOUNDS_CHILD / UNCERTAIN",
      "interaction_type": "ASSESSMENT / PROMPTING / INSTRUCTION_FEEDBACK /
NOISE_CROSSTALK",
      "speech_prominence": "FOREGROUND / BACKGROUND / FAINT",
        "whispering": "YES / NO"
    },
    .
    .
    .
    {
```

```
        "start_timestamp": 5.74, # Speech Segment N
        "end_timestamp": 9.21,
        "speaker_label": "ADULT / CHILD / CHILD_SOUNDS_ADULT /
ADULT_SOUNDS_CHILD / UNCERTAIN",
        "interaction_type": "ASSESSMENT / PROMPTING / INSTRUCTION_FEEDBACK /
NOISE_CROSSTALK",
        "speech_prominence": "FOREGROUND / BACKGROUND / FAINT",
          "whispering": "YES / NO"
    }
}
```