INDIAN INSTITUTE OF TECHNOLOGY PATNA
Department of Computer Science and Engineering
Bihta, Patna - 801106, Bihar, India

EXPRESSION OF INTEREST (EOI)

For Engagement of Startups / Technology Firms / Service Providers

For

Language Data Creation under "COIL-D: Centre of Indian Language Data", a Project under the flagship Bhasini Mission, Ministry of Electronics and Information Technology (MEITY)

Particulars	Details
Date of Issue	October 24, 2026
Last Date for Submission	15 days from the date of issue
Mode of Submission	Email
Contact Email	coild.bhashini@gmail.com
Chief Principal Investigator	Dr. Asif Ekbal, Dept. of Computer Science & Engineering, IIT Patna

This document is for Expression of Interest only and does not constitute a tender or binding contract. IIT Patna reserves the right to accept or reject any or all applications without assigning reasons.

1. Background and Introduction

The Indian Institute of Technology Patna (IIT Patna), through its Department of Computer Science and Engineering, is executing the project titled “COIL-D: Centre of Indian Language Data” - a nationally significant initiative under Mission BHASHINI of the Ministry of Electronics and Information Technology (MeitY), Govt. of India. The project aims at building large-scale, high-quality datasets in Indian languages, building Machine Translation (MT) systems, creating benchmarks and evaluation frameworks.

COIL-D addresses the critical shortage of curated conversational speech and text data required for the development of next-generation Artificial Intelligence (AI) and Natural Language Processing (NLP) systems.

To accelerate data creation at scale, IIT Patna invites Expressions of Interest (EoI) from eligible startups, technology firms, and specialised service providers who have demonstrated capabilities in language data collection, annotation, transcription, speech recording, or related services.

2. Target Languages and Domains

COIL-D aims to build large-scale, high-quality, multi-modal language datasets across a broad range of Indian languages and thematic domains. The following languages and domains are indicative; the project is not limited to these, and additional languages and domains may be included based on mission requirements and emerging priorities under Bhashin, Govt. of India.

2.1 Target Languages

Data creation is required across English and multiple Indian languages spanning major language families. Priority languages include, but are not limited to:

1. **Hindi** (Indo-Aryan): Primary language of the project. Official language of the Union. Spoken by the largest number of people across North and Central India. Data required across all modalities: text, speech (ASR), TTS, and S2ST.
2. **Maithili** (Indo-Aryan (Eastern)): Eighth Schedule language; spoken primarily in Bihar and the Tarai area of Nepal. Rich literary tradition; recognised as a classical language. Significant low-resource gap in speech and NLP datasets.
3. **Santali** (Austro-Asiatic (Munda)): Eighth Schedule language; spoken by the Santhal tribal community across Jharkhand, West Bengal, Odisha, and Bihar. Written in Ol Chiki script. One of the most critically under-resourced languages in Indian NLP; high priority for speech and text data development.
4. **Odia** (Indo-Aryan (Eastern)): Official language of Odisha; Classical Language status conferred in 2014. Rich heritage of literature and oral tradition. Data required across conversational, formal, and domain-specific registers.
5. **Bengali** (Indo-Aryan (Eastern)): Official language of West Bengal; widely spoken in Bihar, Jharkhand, Assam, and Tripura. Large speaker base with significant digital presence. Data required for dialectal and standard varieties.

Note: English and additional Indian languages including Bhojpuri, Magahi, Angikai, Bodo, Dogri, Gondi, Halbi, Kharia, Mundari, and other Eighth Schedule and tribal languages may be included as project requirements evolve. Applicants with capabilities in any Indian language are encouraged to apply.

2.2 Target Domains

Data must be domain-diverse to ensure the resulting datasets are usable across a range of AI and NLP applications. Priority domains include, but are not limited to:

1. **Education:** School and higher education content, NCERT/SCERT materials, pedagogical dialogues, question-answer datasets, and tutoring conversations across subjects and grade levels.
2. **Judiciary:** Legal terminology, court proceedings, judgement summaries, FIR language, rights-awareness content, and public legal communication in Indian languages.
3. **Healthcare:** Doctor-patient dialogues, health advisory content, disease description, medication instructions, maternal and child health, mental health conversations, and public health announcements.
4. **Science & Technology:** Scientific terminology, STEM content, technology news, popular science in Indian languages, and conversational tech-support dialogues.
5. **Tourism:** Cultural heritage descriptions, travel information, hospitality dialogues, place descriptions, and tourism promotion content related to Indian destinations.
6. **Climate & Environment:** Climate awareness, weather advisory, weather alerts, disaster communication, environmental policy, and nature-related vocabulary in regional languages.
7. **Governance & Polity:** Government scheme descriptions, citizen-government interactions, public service announcements, RTI communication, election-related content, and parliamentary language.
8. **Agriculture:** Crop advisory, mandi price information, irrigation, pest management, Kisan helpline dialogues, agricultural news, and rural livelihood content targeting smallholder farming communities.

Note: Additional domains such as Finance & Banking, Transport & Logistics, Sports & Culture, Social Welfare, Defence & Security, and E-Commerce may be included. Applicants are encouraged to indicate their domain expertise in the EOI response. Balanced datasets are preferred.

3. Objective of EOI

The primary objectives of this EOI are:

1. To identify and empanel capable startups and firms for various data creation workstreams under the COIL-D project.
2. To assess the technical capabilities, infrastructure, and prior experience of potential partners.
3. To establish a panel of pre-qualified vendors for phased engagement as data requirements evolve.

4. To encourage participation from Indian startups and MSMEs working in the language technology and AI data space.

Note: This EOI does not guarantee any work order. Empanelled entities will be engaged through separate agreements based on project requirements and satisfactory performance.

4. Scope of Work — Data Creation Workstreams

IIT Patna seeks partners for one or more of the following data creation workstreams. Applicants may apply for any or all categories:

4.1 Source Data Collection & Validation

1. Identification and aggregation of raw source data from diverse domains including education, agriculture, health, legal, governance, and conversational registers.
2. Collection of textual data from copyright-free or rights-cleared sources such as government publications, NCERT materials, open-access corpora, and licensed news archives.
3. Systematic validation of collected data for domain coverage, linguistic quality, speaker diversity, and source authenticity prior to downstream processing.
4. Metadata tagging for each data unit including domain, source type, language variant, geographic region, date of collection, and rights clearance status.
5. Deduplication, normalisation, and cleansing of collected data to remove noise, duplicates, encoding errors, and non-standard characters.
6. **Estimated volume**: As specified in individual work orders; minimum 10,000 sentences per target domain.

4.1.1 Data Acceptance Parameters:

1. **Language & Script:** Data must be authentic Hindi and primarily written in Unicode Devanagari (UTF-8).
2. **Linguistic Quality:** Text should be grammatically correct, meaningful, coherent, and naturally written.
3. **Spelling & Orthography:** Correct spelling, matras, nukta, conjunct characters, punctuation, and word boundaries must be maintained.
4. **Sentence Completeness:** Sentences should be complete and free from truncated, broken, or meaningless content.
5. **Data Cleanliness:** Data must be free from HTML tags, advertisements, unnecessary symbols, OCR errors, encoding issues, and other noise.
6. **Duplicate Content:** Exact and near-duplicate sentences/documents should be identified and removed to ensure data diversity.
7. **Domain & Content Relevance:** Data should conform to the specified domains/topics and contain meaningful and diverse content.
8. **Named Entities & Numerical Information:** Names, locations, organizations, dates, numbers, currency, units, abbreviations, and other special expressions should be accurately preserved.

9. **Metadata & Source Information:** Each dataset should include relevant metadata such as source, domain, document ID, date, language, and collection method, wherever applicable.
10. **Legal Compliance & Human Validation:** The vendor must provide appropriate copyright/licensing declarations, and a representative sample should be validated by qualified native Hindi linguists before final acceptance.
11. **Text Adequacy:** The text should adequately convey the intended meaning, information, and context without omission, distortion, or irrelevant additions. Adequacy should be rated on a 1–5 scale, where 1 = Poor and 5 = Excellent. (1. Poor, 2. Average, 3. Good, 4. Very Good, and 5. Excellent)
12. **Text Fluency:** The text should be grammatically correct, natural, coherent, readable, and appropriate for native Hindi usage. Fluency should be rated on a 1–5 scale, where 1 = Poor and 5 = Excellent. (1. Poor, 2. Average, 3. Good, 4. Very Good, and 5. Excellent)

4.2 Parallel Corpora Creation & Validation

1. Creation of sentence-aligned parallel text corpora between Hindi and one or more Indian languages, including but not limited to Maithili, Bengali, Odia, and Santhali.
2. Translation of source text by qualified bilingual translators with domain expertise; post-editing of machine-translated drafts by certified linguists.
3. Quality validation against accuracy, fluency, and adequacy benchmarks.
4. Domain-balanced corpus construction ensures proportional representation across registers: formal, conversational, technical, and colloquial.
5. **Estimated volume**: 50,000–5,00,000 aligned sentence pairs per language pair per phase.

4.2.1 Data Acceptance Parameters:

1. **Semantic Consistency:** Checks whether the meaning, intent, relationships, and contextual information of the Hindi sentence are consistently preserved in the target translation.
2. **Translation Completeness:** Verifies that all words, phrases, entities, numbers, dates, and important information in the Hindi source are appropriately represented in the target language.
3. **Terminology Accuracy:** Evaluates the correct translation and consistent use of domain-specific, technical, administrative, legal, medical, agricultural, and other specialized terminology.
4. **Named Entity & Numerical Accuracy:** Checks the correct preservation or appropriate transliteration/translation of names, places, organizations, dates, quantities, percentages, measurements, and numerical expressions.
5. **Grammatical & Syntactic Correctness:** Evaluates correct sentence structure, agreement, tense, aspect, case markers, postpositions, gender/number/person, and other target-language grammatical features.
6. **Lexical & Cultural Appropriateness:** Ensures that vocabulary, idioms, expressions, cultural references, and language-specific usages are appropriate and meaningful for native speakers of the target language.
7. **Alignment & Segmentation Quality:** Verifies accurate sentence-level alignment between Hindi and target-language text, including correct segmentation and avoidance of missing, duplicated, merged, or mismatched sentence pairs.
8. **Data Cleanliness & Validation:** Checks for duplicates, irrelevant content, source-target mismatch, untranslated text, excessive code-switching, corrupted characters, formatting issues, and other noise before final corpus acceptance.

9. **Adequacy (1–5):** Measures how accurately the target-language translation conveys the complete meaning and information of the Hindi source without omissions, additions, or distortions. The adequacy should be at least 4.
10. **Fluency (1–5):** Measures the grammatical correctness, naturalness, readability, and linguistic acceptability of the translated sentence in the target Indian language,
11. Recommended Adequacy/Fluency scale: 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, 5 = Excellent..
12. The fluency and adequacy should be at least 4.

4.3 ASR Data Creation (Speech-to-Text)

1. Collection of speech data from diverse native speakers across dialects, ages, genders, and geographic regions.
2. Read speech based on domain-diverse prompts and different sentence types.
3. Spontaneous and conversational speech under controlled and natural acoustic conditions.
4. Coverage of neutral and emotional speech—including happiness, sadness, anger, fear, surprise and disgust—at different intensity levels.
5. Manual transcription in the native script with timestamps and speaker diarisation.
6. Annotation of emotion, intensity, disfluencies, code-switching, noise, and pronunciation variants, where present.
7. Delivery in `.wav`/`.flac` format (16 kHz, 16-bit, mono) with JSON/CSV metadata.
8. **Estimated volume:** 15–100 hours of accepted transcribed audio per language per phase.

4.3.1 Data Acceptance Parameters

1. **Authenticity:** Speech should be human-recorded (representatives from different genders and from different geographical and cultural backgrounds). Participants should be native speakers and proficient in the respective language. Appropriate consent should be obtained from the participants.
2. **Format:** Audio shall be mono, lossless `.wav`/`.flac` at 16 kHz and 16-bit.
3. **Audio quality:** Clean recordings shall have an SNR of at least 20 dB, without clipping, corruption or truncated speech.
4. **Transcription:** Transcripts shall match the spoken content, with audited WER not exceeding 2% for read speech and 3% for spontaneous/noisy speech.
5. **Segmentation:** Utterances shall follow natural boundaries, with timestamps accurate within ±200 ms.
6. **Diarisation:** Multi-speaker turns shall be assigned to the correct speaker; overlapping speech shall be marked.
7. **Emotion annotation:** Each applicable utterance shall contain emotion and intensity labels; blended or changing emotions shall include secondary labels or timestamps.
8. **Emotion validation:** At least two of three trained listeners shall confirm the primary emotion; disagreements shall be adjudicated.
9. **Coverage and metadata:** Approved speaker, domain, dialect, emotion and acoustic-condition distributions shall be maintained with at least 99% metadata completeness.
10. **Batch acceptance:** Critical defects shall have zero tolerance, and the audited major-defect rate shall not exceed 2%.

4.4 TTS Data Creation (Text-to-Speech)

1. Recruitment and recording of approved professionals, including adult male and female speakers and, where required, children, with clear pronunciation representative of the specified language variety.
2. Recording of phonetically and prosodically balanced scripts designed to maximise phoneme coverage, speaking-style diversity, emphasis, pitch variation, and sentence-level intonation.
3. Coverage of neutral and primary emotions-happiness, sadness, anger, fear, surprise, and disgust- with low, medium, and high intensity levels.
4. Inclusion of a controlled subset of naturally blended emotions, such as happiness–surprise, sadness–anger, fear–surprise, and anger–disgust. Selected prompts shall also be recorded across multiple individual emotions for parallel emotional coverage.
5. Annotation of each blended sample with intended primary and secondary emotions, intensity, and perceived emotion distribution based on ratings from at least three trained listeners. Emotion transitions within an utterance shall be time-stamped where applicable.
6. Studio-grade audio capture under consistent recording conditions, with a noise floor below −60 dBFS and an SNR of at least 35 dB.
7. Utterance-level segmentation with precise boundaries. Clipping, false starts, intrusive breath or mouth noise, and editing artefacts shall be removed, while natural breaths shall be retained where necessary.
8. Quality control through listening tests, spectrogram review, prompt-matching checks, emotion-verification tests, and automated audio-quality validation.
9. Delivery as mono, lossless PCM `.wav` files at a minimum of 44.1 kHz and 24-bit, with transcripts, emotion and intensity labels, prosody annotations, speaker metadata, and recording details.
10. **Estimated volume:** 20–50 hours of accepted speech per speaker per phase, with the neutral, single-emotion, and blended-emotion distribution specified in the work order.

4.4.1 Data Acceptance Parameters

1. **Authenticity:** Audio shall be human-recorded by approved, consented speakers; synthetic, cloned, or substitute voices shall be rejected.
2. **Format:** Audio shall be mono PCM .wav, minimum 44.1 kHz/24-bit; lossy or upsampled files are unacceptable.
3. **Audio quality:** Noise floor shall be below −60 dBFS and SNR at least 35 dB, without clipping, distortion or intrusive noise.
4. **Prompt accuracy:** Speech shall exactly match the transcript, with correct pronunciation and no omissions, additions, or false starts.
5. **Segmentation:** Each file shall normally contain one complete utterance of 10–30 seconds and not exceed 30 seconds (these are the tentative values; and the actual thresholds will be decided depending upon the task and domains). Speech shall not be cut, and leading or trailing silence shall remain within prescribed limits.

6. **Coverage compliance:** The accepted corpus shall satisfy the approved language, domain, speaker, phonetic, sentence-type, speaking-style, emotion, and intensity distribution specified in the work order.
7. **Single-emotion validation:** The intended emotion and intensity shall be independently evaluated by at least three trained listeners. At least two listeners must identify the intended emotion, and disputed samples shall be adjudicated or re-recorded.
8. **Blended-emotion validation:** At least two of three listeners must perceive both intended emotions, with the designated primary emotion receiving the higher mean strength. Perceived emotion scores and timestamped transitions, where applicable, shall be supplied.
9. **Metadata completeness:** At least 99.5% of records shall contain utterance ID, speaker ID, language, domain, transcript, emotion, intensity, speaking style, prosody tags, audio specifications, duration, session details and consent reference.
10. **Batch acceptance:** All files shall undergo automated validation, followed by manual review of at least 5% of each batch or 500 utterances, whichever is greater. Naturalness, pronunciation, and expressiveness shall each achieve a mean MOS of at least 4/5; critical defects shall have zero tolerance, and the major-defect rate shall not exceed 2%.

4.5 Speech-to-Speech Data Creation

1. Creation of utterance-aligned parallel speech corpora for Speech-to-Speech Translation between the specified source and target languages.
2. Each source–target pair shall be recorded by the same approved bilingual speaker, based on human translation or interpretation.
3. Delivery of accurate transcripts in the prescribed native scripts for both source- and target-language recordings.
4. Coverage of multiple domains and speaking styles, including narrative, informational, command–response, and spontaneous dialogue.
5. Inclusion of emotion-rich speech with primary emotion, intensity, prosody, and speaking-style annotations.
6. Balanced representation of approved bilingual speakers across gender, age, dialect, and region, with a minimum of 10 speakers per language pair per phase.
7. Delivery as aligned audio pairs with unique pair IDs and JSON/CSV manifests containing transcripts, speaker details, language, domain, emotion, and recording metadata.
8. **Estimated volume:** 30–100 hours of aligned speech pairs per language pair per phase.

4.5.1 Data Acceptance Parameters

1. **Speaker verification:** The speaker ID and verification records shall confirm that both sides of every pair were recorded by the same approved bilingual speaker.
2. **Pair alignment:** Every source recording shall map to exactly one target recording using a unique pair ID. Missing, duplicated, reversed, or mismatched pairs shall be rejected.
3. **Semantic equivalence:** The target speech shall preserve the source meaning, intent, entities, numbers, polarity, and communicative purpose without unjustified additions or omissions.
4. **Transcript accuracy:** Accurate transcripts in the prescribed native scripts shall be provided for both source and target recordings. The audited transcription WER shall not exceed 2%.

5. **Prosody preservation:** Emotion, emphasis, politeness, pauses, speaking style, pitch movement, and speaking rate shall remain functionally equivalent across the two languages.
6. **Emotion-rich data:** Each source–target pair shall be annotated with one primary emotion - neutral, happiness/joy, sadness, anger, fear, surprise, or disgust- and an intensity level (low, medium, or high). Additional affective or speaking-style tags, such as calmness, excitement, frustration, empathy, politeness, urgency, and hesitation, may be specified separately. The source and target recordings shall preserve the same intended emotion and comparable intensity.
7. **Audio specifications:** Audio shall be delivered as mono, lossless PCM `.wav` files at a minimum of 44.1 kHz and 24-bit, using identical technical specifications on both sides.
8. **Audio quality:** Recordings shall have an SNR of at least 30 dB and contain no clipping, truncation, intrusive noise, excessive reverberation, or processing artefacts.
9. **Metadata and authenticity:** Mandatory metadata and valid speaker consent shall accompany every pair. Synthetic, cloned, voice-converted, or cross-speaker audio shall not be accepted.
10. **Human quality evaluation:** Audited pairs shall obtain an average score of at least 4/5 for semantic adequacy, fluency, naturalness, and prosodic/emotional equivalence. Any pair scoring below 3 on any parameter shall be corrected or re-recorded.

4.6 Sentiment Analysis

Sentiment Analysis data comprises labelled text and/or speech samples that enable training and evaluation of models capable of identifying and classifying the affective orientation and opinion expressed in natural language utterances. Partners are invited to collect, annotate, and validate sentiment datasets across target Indian languages and domains.

1. Collection of textual and speech data from diverse sources, including social media posts, product reviews, news commentary, citizen feedback portals, and conversational data — across Hindi and target Indian languages.
2. Sentiment annotation at two levels: **Document-level** (overall polarity of a text), **Sentence-level** (polarity per sentence).
3. Sentiment label taxonomy: Minimum three-class scheme — **Positive, Negative, Neutral**. Intensity in the range of 1-3 (1-lowest and 3-highest) is to be assigned for every positive or negative polarity.
4. Domain-specific sentiment datasets required for: **E-Governance feedback, agricultural advisory reviews, healthcare patient feedback, and tourism review analysis** — aligned with the target domains specified in Section 2.2.
5. Delivery format: **TSV/CSV with UTF-8 encoding**, columns: [ID | Text | Language | Domain | Sentence-level Label | Aspect Sentiment | Annotator IDs | Confidence Score]; accompanied by full annotation guidelines and adjudication logs.
6. Estimated volume: Minimum 10,000 sentences for all the domains as mentioned in Point 4 above.

4.7 Quality Assurance and Validation

Any form of data supplied by the third-party vendor shall be accepted only after the automated and manual quality checks against predefined linguistic, technical, content, metadata, and legal parameters. The data must be authentic Hindi text in Unicode

Devanagari, free from significant duplication, noise, encoding errors, irrelevant content, and truncated sentences. The dataset should satisfy the specified domain and metadata requirements and must have appropriate copyright/licensing clearance. A representative sample of the submitted data shall be manually evaluated by qualified native linguists, and the dataset shall be accepted only if it meets the minimum quality thresholds agreed upon by the Manager/SRA-Language/Respective Authority of the project team. (*Mandatory)

5. Eligibility Criteria

Applicants must satisfy the following eligibility conditions:

5.1 Legal and Organisational Status

1. Registered entity in India — Startup Private Limited, LLP, or Proprietorship.
2. Valid GST registration and PAN.
3. Minimum 1 year of operational history (startups incorporated within 1 year may apply with justification).

5.2 Technical Capability

1. Prior experience in at least one of the workstreams listed in Section 3.
2. Access to data collection infrastructure — recording studios, mobile units, or a network of field collectors.
3. Proprietary or licensed data annotation/transcription tools.
4. Quality management system with measurable accuracy benchmarks.

5.3 Human Resources

1. Team with domain expertise in AI, NLP, Linguistics, Speech technology, or Language services.
2. Access to native language speaker networks (mandatory for Workstreams 3.1 and 3.2).
3. Demonstrated ability to manage distributed data collection teams.

5.4 Data Security and Ethics

1. Commitment to IIT Patna's data governance and privacy policies.
2. Ability to sign a Non-Disclosure Agreement (NDA) and a Data Usage Agreement.
3. No prior history of data misuse, IP violations, or regulatory non-compliance.

6. Expected Deliverables and Quality Standards

Workstream	Primary Deliverable
Speech Collection	Timestamped .wav / .flac files + metadata
Transcription	Respective language script text with timestamps
Text Curation	Cleaned, deduplicated text corpus
Parallel Corpora Creation	Clean data with acceptable quality (prepared from the very scratch)
Sentiment data annotation	High quality sentiment annotated data with acceptable criterion
Validation	Audit report + corrected dataset

All datasets must be delivered in formats compatible with standard AI/ML frameworks (JSON, CSV, TextGrid, ELAN).

7. Indicative Commercial Terms

The following are indicative terms. Final commercial arrangements will be agreed upon during engagement:

Parameter	Details
Payment Basis	Per deliverable unit (per audio hour / per 1000 words / per annotation task)
Payment Schedule	Milestone-based, upon acceptance of quality-verified deliverables
GST	Applicable as per GoI norms; TDS will be deducted at source
Penalty for Non-Quality	Rework at vendor's cost; repeat failures may lead to de-empanelment
Intellectual Property	All datasets and derived works vest exclusively with IIT Patna / funding agency

Note: *Rates will be finalized based on market benchmarks, complexity of task, and negotiation with shortlisted entities after EOI evaluation.*

8. How to Submit — Expression of Interest

Interested entities must submit the following documents:

#	Document / Information Required	Format
1	Company/Organization Profile (max 4 pages)	PDF
2	DPIIT / GST / PAN Registration Certificates	PDF
3	Workstream(s) of interest with brief capability statement	PDF / Word
4	Portfolio of at least 2 relevant past projects with client references	PDF
5	Team composition with CVs of key personnel (max 2 pages each)	PDF
6	Sample deliverable (e.g., 5-minute audio clip + transcript, 500-word annotated/translated text from Hindi to Indian languages)	As applicable
7	Data security and privacy policy / NDA willingness	Declaration on letterhead
8	Indicative pricing per unit	Excel / PDF

8.1 Submission Address

All documents must be submitted via email to:

Email: coild.bhashini@gmail.com

Subject Line: EOI Response — COIL-D Data Services — [Workstream Name] — [Company Name]

Hard Copy (optional): Dr. Asif Ekbal (PI), COIL-D Project, Dept. of CSE, IIT Patna, Bihta, Patna — 801106

Note: IIT Patna reserves the right to empanel multiple vendors per workstream, negotiate terms individually, and issue split work orders based on capacity and performance.

9. Indicative Timeline

Milestone	Indicative Date
Issue of EOI	October 24, 2026

Last date for submission of EOI response	November 8, 2026
Screening, Technical Evaluation Presentations / Demos (if required): ***Only for the shortlisted vendors***	November 23, 2026
Empanelment notification	December 01, 2026 *(Tentative)*
First work orders issued (Phase 1)	December 08, 2026 *(Tentative)*
Works to be completed	February 15, 2027

All timelines are indicative and subject to change at the discretion of IIT Patna.

10. General Terms and Conditions

1. This EOI is not a tender and does not create any legal obligation on IIT Patna to procure services.
2. IIT Patna reserves the right to modify, withdraw, or cancel this EOI at any time without assigning any reason.
3. Responses to this EOI are at the sole cost and risk of the applicant. IIT Patna will not reimburse any costs incurred.
4. All information submitted shall be treated as confidential. IIT Patna will not disclose applicant information to third parties.
5. Intellectual property rights to all datasets, tools, and outputs developed under the project shall vest with IIT Patna and the funding agency.
6. Applicants must disclose any conflicts of interest, including prior affiliations with IIT Patna staff or project team.
7. Sub-contracting of work is not permitted without prior written approval from IIT Patna.
8. Disputes, if any, shall be subject to the jurisdiction of courts in Patna, Bihar.
9. IIT Patna follows Government of India procurement guidelines, and all engagements will be conducted with full transparency and integrity.

11. Contact Information

Role	Contact Details
Project Chief Investigator	Dr. Asif Ekbal, Dept. of CSE, IIT Patna
Project Coordinator / EOI Queries	coild.bhashini@gmail.com, ainlpmliitp@gmail.com
Institute Website	www.iitp.ac.in
Project Reference	COIL-D — Centre of Indian Language Data

For and on behalf of IIT Patna

Dr. Asif Ekbal,
Chief Principal Investigator, COIL-D Project
Department of Computer Science and Engineering,
Indian Institute of Technology Patna-801106