HACKATHON 2025

INSTRUCTIONS FOR SUBMISSION ROUND (Round 2)

1. **Introduction:**
   With the increased thrust on infrastructure development, the construction domain will continue to play a key role in Indian economic development. An integral part of all infrastructure projects is the legally binding construction contract. While construction contracts ensure clarity in risk allocation, parties often enter into an agreement without getting into the fine print in contract documents, later getting into claims and disputes over contract interpretation. One important contract document often the epi-centre of contractual disputes is the "technical specifications." This document contains an employer's technical requirements and is crucial to arrive at the project cost. Research points to the lack of time and expertise to review contract documents as a root cause for an insufficient analysis and understanding of this important document before signing the contract. Rather, a heavy reliance is placed on the experience of the person reviewing the bid documents, which leads to unpleasant financial surprises during project execution. Here, Natural language processing (NLP) tools can help construction firms quickly and efficiently review technical specifications and identify the key material requirements and code provisions so that the tendering team can arrive at the project cost with necessary information rather than mere experience-based bidding. In this hackathon, teams are tasked to develop an NLP-based tool to answer questions about the contract document. The participants must choose appropriate Language Models and use techniques that can develop a custom contract document question-answering tool.

2. **Overall Objective**
   Developing a text extraction tool that can retrieve the details of various materials, relevant codes/standards, and associated tests from the given technical specification document.

3. **Reference Construction Specification Document**
   The reference contract document can be downloaded from https://drive.google.com/file/d/1C0fIvfSzGeHjxi4OmfFGFAEio76O0-rx/view?usp=sharing OR https://shorturl.at/ZhGVO
   This document is a Technical Specification document from a public sector firm. The teams are expected to develop a code and generate an output in the format explained in the next section.

4. **Output Requirements**
   The output should contain the following information in the table format below. Two examples are shown for reference:

Sl. No.	Material Name	Test Name/Reference Code/Standard as per the given document (with reference page number)	Specific Material Type/Material Definition	Any other relevant information

1	*Cement*	*1. Compressive strength test – Page 67, IS 4031 (part 6)* *2. Setting time IS 4031 (Part-5)-1988 (Reaffirm-2014)* *2….* *3….* *4 etc.*	*43 Grade Ordinary Portland*	*The teams can think innovatively and come up with any information here that can help in a better and quicker understanding of the technical specification*
2	*Fine Aggregate*	1. Gradation as per Table 3.1 in Page 67 2. Silt content……	Aggregate most of which passes through 4.75 mm IS sieve	
3	……			
4	And so on			

Note:

1. The extracted information must be from the given reference document only. If the document does not contain information as required, that particular cell in the table should indicate **"No Information Available"**.
2. The above entry shown in the table is only for example. The code should be capable of extracting information regarding **all materials** in the given document. The code must also be capable of receiving and extracting information from any other technical specification document.
3. The teams must also suggest a suitable metric to assess the model's accuracy. However, the final decision on accuracy values will be made per the metric that the organisers will employ during the evaluation.
4. If the team qualifies for Round 3, the judges may give any document, and the qualified teams must be able to extract a similar table from the given document.

5. **The following output files are expected to be submitted for this round by 02/07/2025 23:59 hours. The submission link can be found on the Unstop portal.**

   a. **Output file (.pdf format)**
      - This file should contain the extracted solutions in the table format explained in the 'Output Requirements' (Section 4) above.
      - The file should be submitted in .pdf format.
      - The file's first page should contain your Unstop Team ID/Team Name & all the team members, highlighting the leader.

   b. **Presentation (.pdf format)**
      - This presentation should contain the extracted solutions for the five questions above in the format explained in the 'Output Requirements' (Section 4) above.
      - The Presentation should be limited to 12 Slides (including the title slide). Any team that exceeds this limit will be automatically disqualified.
      - The presentation should be submitted in .pdf format to prevent anyone from using animations.

- The title slide of the presentation should contain your Unstop Team ID & all the names of members of the team highlighting the leader.

c. Zipped version of Folder Containing Code (.pynb Format) & Supporting files (.zip format)

- The code file format should be. pynb and should be able to run in Google Collab independently with the datafile as input in the form of .csv or any other format in the same zipped folder.
- The code should be free from errors and checked before submission. Any code that won't work directly on Google Collab will be rejected.
- Teams must provide hashed-out text in the code sections to explain how the commands are utilised for the problems.

d. 5-minute video (.mp4 format)

- The teams can use any graphics or presentation mode, including the Python code / Slides / Results, etc.
- It is not necessary to show your face in the video.
- The teams are expected to cover five main points in the video as given below:
  - **Explainability** - How does the developed code function for the specific case of question-answering interface?
  - **Output conformance** – How the output confirms the requirements as explained in the section on "Output Requirements."
  - **Novel Understanding** - What insights can be generated and explained from the results?
  - **Innovation –** How does your model exceed expectations beyond what is mentioned under "Output Requirements?"
  - **Accuracy** – How do you demonstrate your model accuracy in the information extracted? How do you establish the output is free from hallucination?

NOTE: If there are file upload issues, please contact the organizers immediately

6. Evaluation criteria (100% split)

- Code quality - 10% (Clear documentation, coding guidelines, proper refactoring)
- Compliance with output requirement (minimum requirements) format – 50%
- Innovation - 10%
- Performance /accuracy of the solution – 20%
- Scalability/potential for large-scale deployment - 10%
- Minimum for qualification to Round 3 – 70%

7. Other important notes

- The decision of the organizers will be final and binding
- Participants should make their own arrangements for APIs/access to Language Models if needed. The organizers will not be providing any API/Language Model access.
- Contact the organizers (contact details available on Unstop) if any clarifications are required.
- Teams cannot use commercially available software.