PS-05: Intelligent Multilingual Document Understanding

1. **General Description**
   a. With the explosion of digital technology, we have amassed a huge collection of documents in various formats. These documents are structured, visually rich, and multilingual—ranging from legal contracts, academic papers to business reports, government forms, and presentation decks, social media posts, and personal screen shots. These documents exist in diverse formats such as Word documents (DOC, DOCX), PDFs, PowerPoint slides (PPT), and scanned images (jpeg, png) and including handwritten documents, often containing mixed scripts (e.g., English-Arabic or Hindi-English, Chinses-English etc).
   b. Traditional OCRs do not extract semantically correct text from various elements present in the documents. The elements may be Table, Image, Maps, Charts. The extracted text is inadvertently mixed/jumbled which is not conducive for downstream AIML related tasks. Moreover, mixed scripts present in document makes the text extraction challenging.
   c. The AI systems need to generate structured outputs from these documents for better indexing and maintaining document layout, enabling better multilingual key word search and boosting performance of various downstream models such as machine translation, vector search, Named Entity Recognition, Retrieval Augmented Generation.

2. **The Challenge**
   a. Multilingual layout-aware document parsing across scripts, formats, and writing directions. The following languages will be present in the documents:
      i. English
      ii. Hindi
      iii. Urdu
      iv. Arabic
      v. Nepalese
      vi. Persian

   b. Accurately extracting structured information from documents while preserving:
      i. Visual hierarchy (headings, sections),
      ii. Semantic grouping (form fields, captions, references),
      iii. Layout fidelity (table structures, image alignment, reading order),
      iv. Embedded elements like charts, plots, maps, and figures.
      v. Representing the extracted content in a standardized, machine-friendly yet human-readable format.
   c. A document may contain the following:
      i. Plain text (headers, paragraphs)
      ii. Table
      iii. Image
      iv. Map
      v. Charts

d. The solution should be able to localize the above components and convert them into natural language text and provide the output in json format with languages identified. The output for the Table, Image, Map, Charts should be the natural language description of the contents. For the text extraction the output should be line wise bbox co-ordinates with the contents and language.

e. The envisaged output from each element present in a generic document mentioned above is indicated in the table below (in json format): -

SNO	Element	Output
1	Table	Table to natural language explain the contents into plain text paragraphs with the bbox co-ordinates
2	Image	Image to natural language description of the contents of the image with bbox co-ordinates
3	Map	Map to natural language description of the contents of the Map With bbox co-ordinates
4	Chart	Chart to natural language description of the contents in the chart With bbox co-ordinates
5	Text	Extracted text with bbox co-ordinates

3. Dataset: The indicative datasets would be used for preparing the first two stages for train/test datasets (not limited too):

DocLayNet
PubLayNet
RvlCdip
ICDAR-MLT 2019
HI-OCR
FUNSD
SROIE

In third stage, organisation specific data would be used.

Metrics For evaluation

Stage	Evaluation Parameters	Remarks
1st Stage	The evaluation metrics: (a) Document Layout: Mean Average Precision (MaP) at bbox threshold >= 0.5	Only Document Layout would be evaluated in first stage
2nd Stage	Result will be evaluated on the following metrics	Document layout, text extraction and

	a) For Document layout – MaP (Mean Average Precision) b) For Text Extraction -CER (Character Error Rate)/ WER (word Error rate) c) BlueRT + BertScore RT for Chart, Map, to Natural Language Text d) T2T-Gen for Table to Natural Language Text e) Language Identification accuracy, precision ,recall	table/map/chart to text in second stage
3rd Stage	Result will be evaluated on the following metrics f) For Document layout – MaP (Mean Average Precision) g) For Text Extraction -CER (Character Error Rate)/ WER(word Error rate) h) BleuRT + BertScore for Chart, Map i) T2T-Gen for Table toNatural Language Text j) Language Identification accuracy, precision, recall	Document layout, text extraction and table/map/chart to text in second stage

4. Dataset arrangement for Stage-1

a. **Training Dataset (upto 10 GB):** train_set.zip. Participant should use this indicative dataset for solution development, in addition to open source datasets of participants choice. It will be released at T0 i.e. 01-Aug-2025.
b. **Mock Dataset (upto 10 GB):** Mock_set.zip. Participant should test their solution on this dataset, but will not have access to the corresponding ground truth during the Challenge. This dataset is for self-assessment and will not be used for evaluation for selection. Participants need to submit the results of their solution on this set. This set will be released T0 + 45 days i.e. 15-Sep-2025 (single dataset consisting of files in different format). A leader board will be published on this Mock Dataset.
c. **Shortlisting Dataset (upto 10 GB):** short_listing_set.zip will be released at 1100h on 04 Nov 2025. Based on the results submitted on this dataset, by 2359h on 05 Nov 2025, 15-20 participants will be shortlisted for offline solution evaluation.
d. **Holdout Test Set (upto 10 GB):** After the Challenge deadline, a private ranking will be computed using this holdout set. This set will be made available during final evaluation post 04-Nov-2025.

5. Input/ Output Instructions

a. Input

The input to the participants would be jpeg/png images of documents. The datasets may also have rotated/blurred/noisy images also.

b. Output

The participants needs to submit one json corresponding to each input image with the details of bbox in [x,y,h,w] and the class of the detected segments(including the rotated one also). The rotated ones may require de-skewing first. The Ground truth bounding are for the de-skewed images. The ground truth bboxes are in HBB format. For the first stage, the classes of the detected segment would be:

{0:"Background",

1: "Text",

2: "Title",

3 : "List",

4: "Table",

5: "Figure""}

6. **Online solution during Stage-1 (Mock Datasets)**
   a. Solutions are expected to be submitted on Thursday from week commencing 15 Sep 2025, on the Mock Dataset.
   b. Leadersboard will be updated every Tuesday**.**
   c. Scores will be computed based on the evaluation metrics as under: -

Category	**Criteria**	**Description**	**%Weight**
Metric Evaluation	mAP Score	Classification and localization	100

   d. Based on performance on the shortlisting datasets top 15-20 participants will be called for an offline evaluation. Scores will be computed based on evaluation metric as above.

7. **<u>Selection of 15-20 participants for offline-evalution in Stage-1</u>**
   a. On 4rd Nov 2025, the details of Shortlisting Dataset will be made available on the website. Results generated on the Shortlisting Dataset will be evaluated for final selection of 15-20 participants. The number

may vary based on the overall performance at the discretion of the Jury for this Problem Statement. Submissions found Incomplete in any manner will not be considered for further processing. The shortlisted participants will be published along with the cut-off score as per the evaluation criteria. Participants individual scores will be shared over the email.

b. Any kind of unfair means be avoided while developing and generating the solution and results, failing which will leads to cancellation of participation for the grand challenge and organisers can call the next participant from leader-board for evaluation.
c. Scores will be computed based on the evaluation metrics indicated below:

Category	**Criteria**	**Description**	**%Weight**
Metric Evaluation	mAP Score	Classification and localization	100

8. Solution Evaluation at the end of Stage-1 Deadline (Holdout Dataset)

a. Shortlisted participants will be asked to demonstrate their solution at IIT Delhi on completion of stage-1 deadline.
b. Participants will be allotted slots in which they need to run their solution on reference data provided by the organizers on given resources with following specifications: -
    i. OS – Ubuntu 24.04 LTS
    ii. CPU – 48+ core
    iii. RAM – 256+ GB
    iv. GPU - A-100, 40/80 GB
    v. Solution Demo Duration: 02 Hours for each selected participant
c. Based on the results from solution demonstration and presentation, final scores will be computed based on Evaluation Metrics as mentioned below:

Category	**Criteria**	**Description**	**% Weight**
Solution Evaluation	mAP Score	Score based on official metric on hidden hold-out test dataset	50
Robustness	Efficiency	Solution Execution time on hold-out test dataset (average per data point)	10

Resource Utilization	Solution Memory Footprint	Memory used by Solution during execution	10
Approach	Methodologies of Solution Development	Start-up need to present Solution development approaches & proposed Architecture	20
Team Capabilities	Technical Capabilities of Start-up Team	Team Composition, Qualifications, Experience and ability to complete the challenge end to end.	10

d. Participants are free to use any language or development framework for the solution.
e. At most top 6 teams will be selected based on final score for Phase-2

9. Evaluation Criteria for Stage-II and Stage-III would be similar as above, only the metrics would change as per 'Metric for evaluation' in Para 3 above. The relative weightages of various parameters would be released before start of that stage. Apart from the languages given in para 2 a. above, a couple of more languages would be released for Stage-2 and Stage-3 participants.

10. **Sessions with Mentors\Experts**
    a. For Stage-1, the organisers plan to meet participants via online meet or email to resolve their doubts, if any. This provision will be made active from 15th Aug 2025 and details regarding interaction will be shared on this website. Kindly keep viewing this website regularly for updates on this.
    b. There will be sessions with Mentors\Experts in Stage-2 and Stage-3 for the willing selected participants to help them in achieving the best solutions.