IndiaAI Intelligent Document Processing (IDP) Challenge

1. Introduction

IndiaAI, an Independent Business Division (IBD) under the Digital India Corporation (DIC) of the Ministry of Electronics and IT (MeitY), is the implementation agency of the IndiaAI Mission, which aims to democratise AI's benefits across all strata of society, bolster India's global leadership in AI, foster technological self-reliance, and ensure ethical and responsible use of AI.

As part of this Mission, the IndiaAI Application Development Initiative (IADI) aims to promote the development, deployment, and adoption of AI applications in critical sectors that have the potential to catalyse large-scale socio-economic transformation.

This Challenge seeks to harness advanced AI/ML technologies including Optical Character Recognition (OCR) and Natural Language Processing (NLP) to optimise public services delivery. Public services like public examinations rely on the authentication and verification of a high volume of text-heavy documents. The existing manual vetting process is constrained by documents of poor quality (e.g., faded text, mixed languages) and inconsistent formats and layouts, resulting in systemic inefficiencies and protracted delays spanning weeks and months.

To address this, the challenge aims to develop an AI/ML solution for the efficient and accurate extraction, summarisation, and verification of critical information from diverse documents. The solution must effectively manage variations such as different layouts, languages, non-standard seals, and poor image quality (e.g., low-resolution, skewed, or distorted scans). Crucially, the solution must extract metadata and summaries from text-heavy documents, like disciplinary proceedings or promotion files, where context is critical, transforming a time-consuming manual review into an actionable, rapid process.

This Challenge offers a platform to create practical, scalable solutions with cross-sectoral and nationwide applicability, enabling seamless deployment and utilisation by various Line Ministries and Central/State government departments.

2. Challenge Process and Participation Details

I. Stage 1 (Application and Virtual Challenge):

a. Participation Requirement:

- Participants may participate as teams as per the eligibility criteria detailed in Section 7.

b. Participation Process:

- All participants are required to utilize the IndiaAI portal to access the application form.
- A team leader will have to individually apply for the Challenge on the IndiaAI portal by clicking the submit link.
- After the initial sign-up, the Team Leader should list all the Team Members under the Management Team and complete all organization details.
- Team Leaders may be required to share the source code required to build, train, and test the submitted AI solution in a private repository. The Team Lead shall paste the link for their GitHub source code in the application form.
- As part of the application stage, participants will be expected to utilise publicly available textual and image-based datasets like certificates, affidavits, notifications, transcripts, mark sheets, disciplinary documents, promotion and diplomas to demonstrate solution performance by specifying key performance metrics, in alignment with the expected output for the Problem Statement. (Ref. Section 3).
- The Team Leader will have to answer all additional questions, including uploading documents, and click 'Submit'.
- Interested applicants can apply within the specified time period from the launch.
- Any edits to the Source Code post the final date of submission will lead to immediate disqualification of the application.

II. Stage 2 (Shortlisting):

a. Participation Requirement:

- Up to top 5 teams may be shortlisted for Stage 2 to test and refine their solutions on a dataset consisting of sample documents and images.
- Shortlisted teams must sign a Non-Disclosure Agreement (NDA) with IndiaAI and partnering institutions in compliance with the current Indian laws. Entry to round 2 will not be provided to shortlisted teams if they fail to sign the NDA on time.

- *IndiaAI reserves the right to modify the number of qualifying solutions to ensure competition and operational efficiency.*

3. Problem Statement

Objective: Develop a high-accuracy, scalable, and secure AI-powered engine leveraging OCR and NLP technologies to perform end-to-end document processing while ensuring high accuracy and cost optimisation.

The solution must efficiently and accurately extract, segment, structure, and analyze critical information from diverse, complex, and multi-format document sources, including certificates, affidavits, transcripts, diplomas, disciplinary proceedings, promotion files, identity cards among others. A core requirement for the engine is its resilience and robustness against common document variance. This includes handling degraded quality inputs such as low resolution scans, faded ink, watermarks, skew/distortion, and variable lighting; adapting to format complexity involving non-standard layouts, varied formats, and complex multi-font content; and addressing linguistic diversity by supporting translation of documents across multiple Indian languages for data extraction, structuring, and verification to English.

While advanced infrastructure, such as vector databases for embeddings and encrypted storage systems to safeguard original documents, is not mandatory in this phase, solutions should be designed to accommodate these requirements in later stages, ensuring scalability, security, and compliance with data-handling standards. Crucially, any extracted data translated into English may be cross-verified against the corresponding application forms to ensure integrity in the later stages.

The solution must include the following capabilities:

1. **Data Extraction:** Extraction of text, tables, hyperlinks, embedded data, and other key fields from both structured (tables, forms) and unstructured (text blocks, notes) content across all document types.

2. **Document Segmentation**: Segment documents into logical, header-based sections with appropriate metadata tagging, enabling structured access, navigation, and efficient retrieval.

3. **Structured Output & Summarization:** Generate structured, usable outputs, including concise and verifiable summaries of a document's content, focusing on critical information. For cross-verification purposes, an explainable output detailing the alignment with the source document or application form may be provisioned in the subsequent stages of the solution development.

The technical evaluation will prioritize the solution's performance accuracy, cost efficiency, and demonstrated ability to process documents with diverse quality and formats.

4. Application Requirements

I. Solution Code and Documentation (GitHub):

- Team Leaders may be required to share the source code required to build, train, and test the submitted AI solution in a private repository. The Team Lead shall paste the link for their GitHub source code in the application form.
- Explanation of the key methodology and steps taken in solution development.
- Steps to grant access to your GitHub repository:
    - Go to the main page of your GitHub repository.
    - Click on the 'Settings' tab in the menu bar.
    - In the left sidebar, select 'Collaborators'.
    - Under the 'Manage Access' section, click on 'Add people'.
    - In the text field, search for 'indiaaihackathon25' and add it as a collaborator with read access.

II. Application Form:

- Description of the solution, approach to addressing problem statements, and core AI technologies used. Provide a summary of document completeness or integrity challenges. Present essential metrics (OCR accuracy, segmentation performance, etc.). Cost efficiency of proposed OCR tool across different document types.

- Uniqueness and novelty of the solution in optimising costs along with its suitability for use cases at hand. Replicability and scalability across public and private sectors.

- AI solution details including information about architecture and solution design, third party integration, data utilised for training and validation (including data provenance, coverage and size), details of solution evaluation and outcomes. Please outline solution monitoring and enhancement strategy, including areas that require further refinement and process adopted to integrate improvements.

5. Evaluation Process

The evaluation process for the Challenge will be overseen by a distinguished panel of jury members, comprising subject matter experts in machine learning, data science, alongside domain specialists. The jury would rigorously assess each submission based on predefined criteria to ensure a fair and comprehensive evaluation. The evaluation will ensure equitable weightage is given to both the Technical and General parameters.

Stage 1

- **Initial Screening**: Submissions would undergo an initial screening to ensure compliance with submission guidelines and solution functionality.
- **Technical Evaluation**: The jury would conduct a detailed technical evaluation of the solutions.

Stage 2

- Up to 5 teams may be shortlisted based on the initial evaluation. Each shortlisted team will receive INR 5 lakhs to refine their solutions and submit results on the shared data within the stipulated processing period and requested format.
- Entry to the second stage and access to and use of sample data will require shortlisted applicants to sign a Non-Disclosure Agreement (NDA) in compliance with the current Indian laws.

Stage 3

- Following the assessment of the results submitted by the shortlisted teams, up to 2 teams may secure a chance to get a two-year work order worth up to INR 1 crore to deploy their solution for use by the Government of India and its associated entities to ensure robust cross-verification of results.
- The jury's decision would be final and binding.

IndiaAI and the jury reserve the right to modify the number of qualifying solutions at any stage to ensure competition and operational efficiency.

6. Opportunity for Applicants

- **Opportunity to Build for the Nation**: Contribute to developing innovative solutions that address critical challenges faced by the country, making a direct impact on society.
- **National Recognition**: Gain visibility and recognition from government officials, industry leaders, and peers for your contributions and innovative ideas. Top solutions may be listed on AIKosh.
- **Networking Opportunities**: Connect with like-minded innovators, potential collaborators, and key stakeholders in the tech and innovation ecosystem.

- **Exposure to Real-world Challenges:** Work on pressing issues faced by the nation, providing practical experience and a deep understanding of real-world problems.
- **Support for Implementation**: The winning solution will get potential support in scaling and implementing the solution at a national level, bringing your ideas to life.

7. Eligibility

- **Indian Company**: Indian company registered under the Companies Act, 2013. An Indian company must have 51% or more shareholding by Indian citizens or persons of Indian origin.
- **Start-up**: Start-up as defined in the latest notification by the Department for Promotion of Industry and Internal Trade (DPIIT), accessible at Startup India.

The participating entities must provide proof of a proprietary solution submitting verifiable documentation (e.g., technical specifications, necessary certifications, IP and patents). Additionally, entities must demonstrate development and deployment experience of relevant solutions with previous engagements in the private and public sector.

8. Timeline

#	Activity	Timeline
1	Launch Date	27-11-2025
2	Last Date for Online Submission	23-12-2025
3	Announcement of Results of First Round	TBC
4	Round 2 (On-premise, New Delhi)	TBC
5	Announcement of Winning Entity and pilot	TBC

9. Intellectual Property Rights

All Intellectual Property Rights (IPR) will belong to the solution owner participating in the Challenge. IndiaAI and partnering institutions shall have a non-exclusive, royalty-free, perpetual license to use the awarded AI solution including all Intellectual Property Rights arising out of its use, and the solution owner shall be deemed to have given a No Objection Certificate (NOC) for the same and shall also remain bound by the terms of a Non-Disclosure Agreement (NDA) with respect to such work.

10. Terms and Conditions

a. All participants must meet the outlined eligibility criteria (Section 7) and belong to legally registered entities as of launch date of the Challenge.

b. The award from this initiative can only be used by the participating teams for the purpose of AI solution development.

c. Winning entities will retain the rights to the solution/product developed subject to the intellectual property rights outlined in this document (Section 9).

d. IndiaAI may host up to top 5 solutions refined under Stage 2 on AIKosh for public use, subject to applicant consent and compliance with AIKosh guidelines.

e. The participants will ensure code is free from viruses and malware. The participants will not use this Challenge to do anything unlawful, misleading, malicious, or discriminatory.

f. The solutions must not violate/breach/copy any copyrighted or patented concepts in the AI market.

g. The solutions must not violate any data protection and governance regulations and policies.

h. The solutions must be in adherence with related cybersecurity standards and guidelines of the Government of India.

i. Solutions must adhere to ethical principles and guidelines for the development, deployment and use of AI technologies, including fairness, transparency, accountability, and non-discrimination.

j. The developed solution/product will be deployed in the chosen Cloud Environment and used for Union/State/UT government entities.

k. Any new enhancements, features or innovation should be released on the chosen Cloud Environment.

l. The winning entities shall receive a work order of a fixed amount to support the solution development and deployment for at least two years from the go-live period. The support includes manpower for end-to-end development, deployment, maintenance, and bug fixing across the entire application.

m. The winning entities shall submit progress-cum-achievement reports at quarterly intervals on the progress made on all aspects of the project, including expenditure incurred on various approved items during the two-year period. The scope of work, payment terms, milestones, and other conditions will be as agreed between IndiaAI, partnering institutions

and the winning entities and it shall comply with the General Financial Rules of the Government of India.

n. The winning entities are not allowed to entrust the implementation of this project for which the award is received to another institution, and to divert the award received from IndiaAI as assistance to the latter institution.

o. The winning entities should not enter into collaboration with a foreign party (individual/academic institution/industry) in execution of this project without prior approval of IndiaAI.

p. The winning entities are free to market the product to any entity outside the Union/State/UT Government Organisations of India.

q. In case of any dispute on any other matter related to the project during the course of its implementation, the decision of the CEO, IndiaAI shall be final and binding on the winning entities.

r. IndiaAI reserves the right to modify the terms and conditions of this challenge for operational feasibility and compliance to rules and regulations of the Government of India.

s. By participating in this Challenge, the winning entities understand and undertake the above commitments and agree to the terms and conditions.

11. Plagiarism and Ethics

a. Participants are expected to uphold the highest standards of ethics and integrity throughout the Challenge.

b. All work submitted must be original and developed by the participant or their team.

c. Plagiarism, or the use of someone else's work without proper attribution, is strictly prohibited and would result in immediate disqualification.

d. Participants must ensure that their solutions are proprietary and not copied from existing projects or code repositories.

e. Moreover, the use of any external resources or pre-trained models should be clearly cited, and proper permissions should be obtained where necessary. Adherence to these ethical guidelines ensures a fair and competitive environment for all participants.

f. By registering for this Challenge, participants are giving an undertaking to adhere to all plagiarism and ethical guidelines set forth by the IndiaAI.

Annexure I: Evaluation Parameters

I. General

	Parameter	Description
1	**Approach Towards Problem Solving**	Problem Formulation, Product Idea, Methods adopted, Simplicity of Final Solution, Uniqueness of Idea, Novelty of Approach, Evaluation Approach.
2	**Solution Technical Feasibility**	Product features, Demonstrated evidence of performance, Scalability, Interoperability, Enhancement & Expansion, Underlying Technology Components & Stack and Future-readiness, and System Integration Plan.
3	**Product Cost**	Potential Cost to Build, Deploy and Maintain Product for two years from go-live.
4	**Team Ability & Culture**	Prior Experience of the entity in developing and deploying similar solutions in private and public domains, Team Leader's Effectiveness (i.e. Understanding of subject matter, Ability to guide, Ability to present idea), Ability to scale up and market the product, Growth Potential of Organisation.
5	**Adherence to Responsible AI Principles**	Safety and Reliability, Equality, Inclusivity and Non-discrimination, Privacy and Security, Transparency, Accountability, Protection and Reinforcement of positive human values.
6	**Adherence to Data Policies and Cyber Security Guidelines**	Adherence to applicable Government of India policies, guidelines, regulations on Data Governance and Cyber Security. Regulatory compliance.

II. Technical

	Parameter	Description
1.	**Data preparation**	Participant has • Ensured data coverage, size, and quality • Performed appropriate exploratory data analysis, to gain a comprehensive understanding of the dataset, including statistical properties, missing values, outliers etc.

		• Appropriate data processing pipeline implemented to prepare data to be ready for modeling • Advanced data analysis and processing to inform modeling approaches as relevant • Split the data appropriately to define training and test sets, with holdout validation sets optionally • Integrated data management and governance policy
2.	**Solution Building**	• Appropriate baselines identified and evaluated • Appropriate strategy defined for modeling and model/hyperparameter selection, using data splits defined earlier • A reasonable number and variety of different AI/ML technologies are attempted, and the best one is chosen based on key performance metrics, demonstrating improvement over baselines. • Feasible strategy for integration to be documented
3.	**Solution Evaluation**	• Solution evaluation is conducted using an appropriate metric (s), on the entire dataset and relevant cohorts. • Evaluation to consider model performance, but also other aspects such as robustness, fairness, efficiency, interpretability etc. as relevant • Integrated Solution monitoring and enhancement strategy.
4.	**Code readability and conciseness**	Efficient, concise code is written. The code is well documented, and the model output analysis is explained in the report format with findings from the dataset.
5.	**Technical Robustness**	Applicants may refer to the below parameters to assess and share solution performance as part of the application process. The metrics should be used for the appropriate task at hand, and measured on the overall set as well as cohorts : Data Extraction:

		- Character Error Rate (CER) on standard public benchmarks for scanned documents, such as the ICDAR 2019 ArT dataset or the SROIE dataset, or equivalent. - Key Information Extraction - Strict, entity-level F1-score on the FUNSD (Form Understanding in Noisy Scanned Documents) benchmark, or equivalent. - Extraction Accuracy - % of correctly extracted text, tables, hyperlinks, and numerical data vs. ground truth - Latency - Average processing time per page/document - Configurability - Ability to adapt compliance tests - Scalability - Performance with increased volume or new data source Document Segmentation: - Mean Intersection over Union (mIoU) for segmenting document regions (e.g., tables, text, figures) on the PubLayNet or DocLayNet benchmarks or equivalent. Generation of Structured Output like Flagging Outcomes & Summarization: - ROUGE-1, ROUGE-2, and ROUGE-L scores on a standard summarisation benchmark like CNN/DailyMail/Xsum or equivalent. Also, BERT Score. - Macro-F1 and Matthews Correlation Coefficient (MCC). - Provide confusion matrix, a table providing a detailed breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Additional Criteria: Any other metrics as agreed upon by jury members.

Annexure II: Application Form - IndiaAI Intelligent Document Processing (IDP) Challenge

Section 1: Organisation Information

1. Applicant Type (Startup/Company)*:
2. Organisation Information*:
    - Name of Organisation:
    - Registration Number:
    - Date of Incorporation:
    - Number of Employees:
    - Core Function of Organisation:
    - Address:
    - Website:
3. PoC Information*:
    - Full Name:
    - Designation:
    - Core Expertise Areas:
    - Email Address:
    - Phone Number:
    - LinkedIn Profile:
4. Organisation Member Information: List details of members to be involved in this challenges
    - Name
    - Role
    - Email
    - LinkedIn Profile
5. Prior Experience in relevant project implementation and research work*:
   Describe relevant AI solutions designed and developed, technologies used, and outcomes, relevant publications or patents, etc. (max 200 words).
6. Experience Collaborating with Government Entities for relevant use cases*:
   Specify partners, nature of engagement, and key results achieved (max 200 words).

Section 2: Project Proposal

1. Approach to Addressing the Problem Statement*: Explain how your solution utilises AI for:
    a) Data Extraction
    b) Document Segmentation
    c) Generation of Structured Output & Summarization

   (Max 300 words)

2. Description of AI Solution*: Provide a comprehensive overview of your AI solution, including:
    - Functionality
    - Features
    - Core AI technologies used
    - Training and validation data used, highlighting data provenance, coverage, size and quality
    - Process and strategies adopted for data preparation, AI/ML technologies selection, training, hyper-parameter tuning, refinement, solution monitoring and enhancement
    - Solution replicability across multiple sectors for relevant use cases *(Max 300 words)*

3. Upload Solution Architecture Diagram*

4. Upload technical performance metrics measured (for each task such as Data Extraction, Document Segmentation and Generation of Structured Output & Summarization based on standard public benchmarks) the methodologies used for measurement, and the outcomes in the below format. Please refer to the 'Technical Robustness' section of the Evaluation Parameters in the Schema Document. *(In the format given in Annexure III, Part A)*

5. Proprietary Solution*:
    - Is the AI solution developed in-house (not based on third-party pre-trained models)?
        - ☐ Yes ☐ No
    - If Yes, provide:
        - Details of a proprietary technology base
        - If the solution is developed on open-source models, share details and customisation approach
        - Details of proprietary data utilised for training and validation, along with explicit confirmation and evidence of adherence to all relevant Indian laws and compliance standards
    - If No, provide:
        - Names of the third-party models or components used, share specific licensing agreements that govern their use
        - Refinement approach
        - Data sourcing, coverage and validation approach *(max 100 words)*

6. Data Governance and Security*: Describe how data collection, confidentiality, encryption, storage, access control, retention and removal

will be implemented. Include measures taken to ensure compliance with relevant regulations and standards for data privacy and security measures *(max 100 words)*

7. Scalability and Integration Readiness*
   Describe deployment mode, integration compatibility, offline operability, and future expansion capability. *(max 100 words)*

8. Compliance with Responsible AI Principles*
   Describe how the solution adheres to principles of fairness and transparency and adopts measures for solution interpretability, auditability, inclusivity and fairness. *(max 100 words)*

9. GitHub Link:

Section 3: Supporting Documents (Upload)

1. Certificate of Incorporation/Legal Registration*
2. Certificate of Recognition (for Startups)
3. Technical Documentation/Proof of IPs/Patents, if any
4. Ethics/Regulatory Clearance (if secured)
5. Solution demo video (2-3 minutes)*
6. Any additional documentation to strengthen the proposal such as solution accuracy proof etc.
7. Estimation of Initial manpower cost for a two-year deployment period after stage 2. Estimation of other Potential costs to build, deploy and maintain the solution, along with the System Integration plan.* *(In the format given in Annexure III, Part B)*

Section 4: Declaration

Declaration by Team Leader:

- I/We declare that all the information provided in this application is true and complete to the best of our knowledge. I hereby also declare that I am authorised by my company/startup to participate in this Challenge, sign all the documents and agreements related to the Challenge and to commit resources.
- I/We confirm that we will abide by the conditions mentioned in the Challenge document in full and without any deviation.
- I/We shall observe confidentiality of all the information passed on to us in the course of the Challenge and shall not use the information for any other purpose than the current Challenge.

- I/We confirm that we have not been blacklisted/banned in the last three years by any State/Central Government organisations/Firms / Institutions/ Central PSU / PSE.

☐ I accept

Annexure III: Template for Solution Technical Performance Matrix and Cost for Development, Deployment and Maintenance and Support

A. **Template for Solution Technical Performance Matrix**

a. Data Extraction

Metric	Definition	Value/Benchmark	Methodology Used
Metric 1: Character Error Rate (CER)	The percentage of incorrectly recognized characters (substitutions, insertions, deletions) out of the total number of characters in the ground truth text. A lower CER is better.	ICDAR 2019 ArT dataset or SROIE dataset, or equivalent	
Metric 2: Key Information Extraction (Strict, entity-level F1-score)	A metric that combines Precision and Recall to measure the accuracy of extracting specific entities (e.g., "Total Amount," "Vendor Name"). The F1-score is the harmonic mean of both, providing a single score for Solution performance.	FUNSD (Form Understanding in Noisy Scanned Documents) benchmark, or equivalent	
Metric 3: OCR Operational Cost	Cost per page character recognition		
Any other Metric			

b. Documentation Segmentation Metrics

Metric	Definition	Value/Benchmark	Methodology Used
Metric 1: Mean Intersection over Union (mIoU)	For segmenting document regions (e.g., tables, text, figures)	PubLayNet or DocLayNet benchmarks or equivalent	

Any other Metric			

c. Generation of Structured Output & Summarization

Metric	Definition	Value/Benchmark	Methodology Used
Metric 1: ROUGE-1, ROUGE-2, ROUGE-L	Measures the overlap of unigrams (individual words), bigrams (pairs of consecutive words), and Longest Common Subsequence (LCS) between the generated summary and a set of reference summaries.	CNN/DailyMail/Xsum or equivalent	
Metric 2: BERT Score	An automatic evaluation metric that computes the similarity between tokens in a candidate summary and a reference summary using BERT embeddings, capturing semantic similarity beyond simple word overlap.	CNN/DailyMail/Xsum or equivalent	
Any other Metric			

B. **Template for Cost for Development, Deployment and Maintenance and Support**

Initial cost and Manpower cost for further development, deployment, maintenance and support for two years from go-live of the solution selected as per IndiaAI Intelligent Document Processing (IDP) Challenge.

(All in INR, excluding GST, and including any overheads, employee benefits, other taxes and levies)

A. Initial Cost

Sl. No.	Cost Component	Total rate	Impact

B. Cost for further enhancements, operations, maintenance and support for two years from go-live (including infrastructural, human resource, security compliance, licensing and others)

Sl. No.	Cost Component	Monthly Rate	Total rate for two years