

I. Problem Statement: (Stage 1) - AI Engine for Compliance Gaps Objective: AI-powered engine capable of extracting text, tables, hyperlinks, embedded data, and financial data from multi-format document sources in scanned and digital versions, segmenting them into logical, header-based sections and validating each section's completeness, integrity and compliance against a pre-defined framework. The solution should structure document metadata and support search, analytics, visualisation and indexing of source unstructured data to ensure efficient retrieval and referenceability. 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.

#### Expected Output

i. Compliance Validation Report Generator: A generic Engine configurable for testing compliance across large volumes of structured or unstructured datasets with a corresponding set of regulations. The Engine should first extract and structure text, numerical and financial data from scanned and digital multi-format documents. It should segment documents into logical, header-based sections with appropriate metadata tagging, enabling structured access and navigation.

The Engine should be able to map the rules and regulations against the documents received providing an explainable output indicating the dataset's compliance and non-compliance with each provision of the rules and regulations. Examples of such rules/regulations/frameworks include Indian Accounting Standards and associated reference and educational materials, SEBI disclosure requirements, RBI Disclosure Norms for Financial Institutions, SEBI ESG Framework, among others<sup>1</sup>. Further, examples of datasets that reflect compliance with these rules/regulations include financial statements, statutory forms, XBRL filings, stock exchange filings,

and so on. The Engine should structure the regulations and datasets or guide the user in structuring them using built-in tools. ii. Project Report: A structured document summarising the approach, methodologies, datasets used, validation results, and challenges encountered during the development of the AI Validator Tool.

#### Submissions

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 & Summarisation (Max 300 words)

2. Description of AI Solution\*: Provide a comprehensive overview of your AI solution, including:  
o Functionality o Features o Core AI technologies used o Training and validation data used, highlighting data provenance, coverage, size and quality o Process and strategies adopted for

data preparation, AI/ML technologies selection, training, hyper-parameter tuning, refinement, solution monitoring and enhancement o 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 & Summarisation based on standard public benchmarks), the methodologies used for measurement, and the outcomes in the format provided in Annexure III, Part A of the Submission Guidelines.\* Please refer to the 'Technical Robustness' section of the Evaluation Parameters in the Schema Document.

5. Proprietary Solution\*: o Is the AI solution developed in-house (not based on third-party pre-trained models)? □ Yes □ No o If Yes, provide: ▪ Details of a proprietary technology base (max 100 words) ▪ If the solution is developed on open-source models, share details and customisation approach (max 100 words) ▪ 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. (max 100 words)

If No, provide: ▪ Names of the third-party models or components used, share specific licensing agreements that govern their use (max 100 words) ▪ refinement approach (max 100 words) ▪ 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)

## 1. Data Extraction

Metric	Definition	Value/Benchmark	Methodology Used
Metric 1: Character Error Rate (CER)	The percentage of incorrectly recognised 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			

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			

### 3. Generation of Structured Output & Summarisation

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	
-------------------------	---	-------------------------------------	--