*IFC_2.0_Q1_*Improve Multilingual OCR Accuracy

Theme	Intelligent Data Processing & Enrichment
Initiative	Improve Multilingual OCR Accuracy
Document ID	PRD-IFC-T2.1-V1.0
Owner	Anuj S.
Team	Data Science Team, Engineering Team, QA Team
Summary	Enhance the accuracy of Optical Character Recognition (OCR) for English, Hindi, and Mandarin in scanned documents, and improve the user's ability to correct OCR errors.
Status	Discovery / Delivery
Next Milestone	OCR Model Training Completion (Q2 2026)

Requirement Definition

Objective

To achieve higher accuracy in converting scanned documents into machine-readable text across supported languages (English, Hindi, Mandarin) and to reduce the manual effort required for OCR error correction, thereby improving the quality of unstructured data for downstream processing and analysis. This aligns with the company goal of "Actionable Intelligence Generation" and "Operational Efficiency."

Background context

Currently, the IFC system ingests scanned documents, but the accuracy of OCR, particularly for non-English languages, can be inconsistent. This leads to a significant amount of manual review and correction by data reviewers, slowing down the overall data processing pipeline and impacting the reliability of search and text analytics. Improving OCR accuracy is a foundational step to unlock the full potential of text analytics and LLM capabilities.

Problems

- **Inaccurate Text Extraction:** OCR errors lead to unreliable search results and poor input for text analytics.
- **High Manual Correction Effort:** Data reviewers spend excessive time correcting OCR errors.

- **Limited Multilingual Support:** Current OCR performance for Hindi and Mandarin is suboptimal.

Constraints

- **COTS Hardware:** Reliance on existing hardware infrastructure for OCR processing.
- **External OCR Engine:** Integration with a specific third-party OCR engine (e.g., Tesseract or a selected commercial SDK).
- **Performance:** OCR processing must not become a bottleneck in the ingestion pipeline.
- **Data Availability:** Availability of sufficient, diverse, and labeled scanned document datasets for model training and testing.

Solution briefing

Our approach is to fine-tune or upgrade the existing OCR engine and implement robust pre-processing techniques for scanned images. We will also enhance the OCR error review interface to make manual corrections more efficient. The focus will be on improving character recognition accuracy for English, Hindi, and Mandarin.

Functional requirements (User Stories with Acceptance Criteria)

- **User Story 3.2.1:** As a System, when a scanned document is ingested, I want to automatically perform OCR on it, so that its text content can be extracted and made searchable. (Ref. FR-OCR-001.1)
    - **Acceptance Criteria:**
        - **AC 3.2.1.1:** GIVEN a scanned document (e.g., JPEG, TIFF, PDF image-only) is successfully ingested via FR-ING-001.4, WHEN the ingestion process completes, THEN the system SHALL automatically trigger the OCR module for that document.
        - **AC 3.2.1.2:** GIVEN OCR processing is complete, THEN the extracted text SHALL be stored in the Big Data Repository (FR-BDR-001) and linked to the original scanned document.
        - **AC 3.2.1.3:** GIVEN OCR processing is complete, THEN the extracted text SHALL be indexed for search (FR-SRCH-001.1).

- **User Story 3.2.2:** As a System, I want to accurately recognize and extract text in English from scanned documents, so that English content is reliably available for analysis. (Ref. FR-OCR-001.2)
    - **Acceptance Criteria:**
        - **AC 3.2.2.1:** GIVEN a clear, standard English scanned document (e.g., 300 DPI, clean font), WHEN OCR is performed, THEN the OCR accuracy SHALL be greater than 95% (character accuracy rate).
        - **AC 3.2.2.2:** GIVEN a standard English scanned document with common formatting (e.g., tables, bullet points), WHEN OCR is performed, THEN the extracted text SHALL retain basic structural elements (e.g., line breaks, paragraph separation, table cell separation).

- **User Story 3.2.3:** As a System, I want to accurately recognize and extract text in Hindi from scanned documents, so that Hindi content is reliably available for analysis. (Ref. FR-OCR-001.3)
    - **Acceptance Criteria:**
        - **AC 3.2.3.1:** GIVEN a clear, standard Hindi scanned document (e.g., 300 DPI, common Hindi font), WHEN OCR is performed, THEN the OCR accuracy SHALL be greater than 90% (character accuracy rate).
        - **AC 3.2.3.2:** GIVEN a standard Hindi scanned document with common formatting, WHEN OCR is performed, THEN the extracted text SHALL retain basic structural elements.

- **User Story 3.2.4:** As a System, I want to accurately recognize and extract text in Mandarin from scanned documents, so that Mandarin content is reliably available for analysis. (Ref. FR-OCR-001.4)
    - **Acceptance Criteria:**
        - **AC 3.2.4.1:** GIVEN a clear, standard Mandarin scanned document (e.g., 300 DPI, common Mandarin font), WHEN OCR is performed, THEN the OCR accuracy SHALL be greater than 90% (character accuracy rate).
        - **AC 3.2.4.2:** GIVEN a standard Mandarin scanned document with common formatting, WHEN OCR is performed, THEN the extracted text SHALL retain basic structural elements.

- **User Story 3.2.5:** As a Data Reviewer, I want to be able to review and correct any errors in the OCR-extracted text, so that the final text content is accurate. (Ref. FR-OCR-001.5)
    - **Acceptance Criteria:**
        - **AC 3.2.5.1:** GIVEN a scanned document with OCR-extracted text, WHEN a Data Reviewer accesses the document in the review interface, THEN the interface SHALL display the original image side-by-side with the editable extracted text.
        - **AC 3.2.5.2:** GIVEN the side-by-side view, WHEN the Data Reviewer makes changes to the extracted text and saves, THEN the corrected text SHALL be updated in the Big Data Repository and re-indexed for search.
        - **AC 3.2.5.3:** GIVEN the side-by-side view, WHEN the Data Reviewer hovers over a word in the extracted text, THEN the corresponding word in the original image SHALL be highlighted (and vice-versa, if technically feasible, for context).

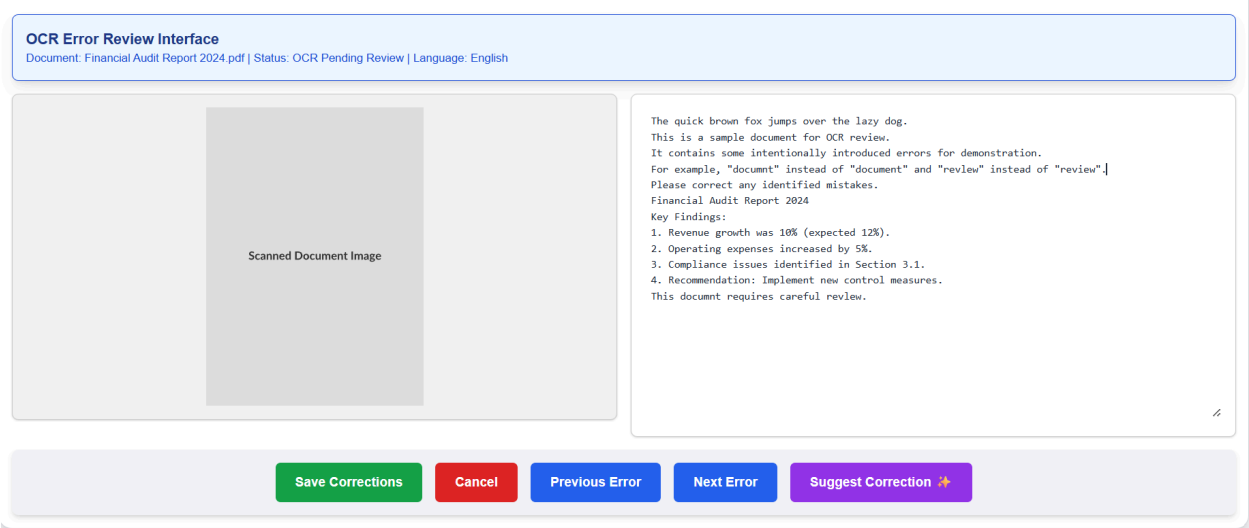
Non-functional requirements

- **NFR-OCR-PER-001:** OCR processing for a standard A4 page (300 DPI, moderate text density) shall complete within 5 seconds.
- **NFR-OCR-REL-001:** The OCR process shall gracefully handle low-quality images (e.g., blurry, skewed) by flagging them for manual review or indicating low confidence, rather than failing the entire ingestion.
- **NFR-OCR-SCL-001:** The OCR module shall be able to process 100 A4 pages concurrently without significant performance degradation.

Designs

Wireframes, mockups and data flow diagrams as they are developed, showing what the feature will look like, and how users will interact with it.

Design view Link: https://g.co/gemini/share/c3e84c1113b0



Reference

- SRS-IFC-V1.0, Section 3.1.2 Multilingual OCR (FR-OCR-001)
- IDD-IFC-V1.0, Screen: Document/Record View (with TA Insights) - for OCR review interface

OKR metric

- **Objective:** Improve OCR accuracy and reduce manual correction effort.
- **Key Results:**
  - Achieve >95% OCR accuracy for English, >90% for Hindi and Mandarin on clear scanned documents by end of Q2 2026.
  - Reduce average OCR review/correction time by 25% by end of Q2 2026.

Stakeholders

- **Responsible:** Engineering Team (OCR Integration), Data Science Team (OCR Model Tuning)
- **Accountable:** Product Manager
- **Consulted:** Data Reviewers, Data Administrators, QA Team
- **Informed:** Project Management, End-Users (Data Analysts, Clerks)

Risks

- **R-OCR-001: Accuracy Degradation for Complex Documents:** OCR accuracy may drop significantly for very complex layouts, handwritten text, or very low-quality scans.
    - **Mitigation:** Focus on "clear scanned documents" as per ACs. Implement confidence scoring. Provide clear guidelines on scan quality.
- **R-OCR-002: Performance Bottleneck:** OCR processing could slow down the ingestion pipeline for large volumes.
    - **Mitigation:** Utilize asynchronous processing. Scale OCR workers/instances. Optimize hardware utilization.
- **R-OCR-003: Multilingual Model Availability:** Difficulty in acquiring or fine-tuning high-quality OCR models for Hindi/Mandarin.
    - **Mitigation:** Research and evaluate multiple commercial/open-source options. Allocate dedicated data science resources for model training.

Decision log

- **Question:** Which OCR engine to prioritize for multilingual support?
    - **Decision:** Start with Tesseract for initial integration due to open-source flexibility. Evaluate commercial alternatives if accuracy targets are not met after initial tuning.
    - **Decider:** Data Science Lead, Engineering Lead
    - **Date:** 2025-05-15

Launch Readiness

Testing plan

- **Unit Testing:** For OCR integration components and pre-processing logic.
- **Integration Testing:** Verify seamless integration with ingestion pipelines and BDR.
- **Functional Testing:** Execute test cases covering all user stories and acceptance criteria, including specific test sets for English, Hindi, and Mandarin documents (clear, slightly degraded, mixed content).
- **Performance Testing:** Measure OCR processing time for various document sizes and concurrent loads.
- **Accuracy Testing:** Develop automated and manual methods to measure character accuracy rate (CAR) and word accuracy rate (WAR) against ground truth data for each language.
- **User Acceptance Testing (UAT):** Involve Data Reviewers to validate the OCR correction interface and overall accuracy.

Tracking & analytics

- **OCR Accuracy Dashboard:** Track CAR/WAR metrics per language over time.
- **Manual Correction Time:** Monitor average time spent by reviewers on correcting OCR errors.
- **Ingestion Pipeline Metrics:** Track OCR processing time per document and overall

throughput.

Marketing plan

- Internal communication to all users highlighting improved search accuracy and reduced manual effort due to better OCR.
- Training sessions for Data Reviewers on the new OCR correction interface.

Customer service plan

- Brief customer service team on common OCR issues and troubleshooting steps.
- Provide FAQs on expected OCR accuracy and how to report issues.

Legal checks

- No specific legal implications beyond existing data handling policies.

FAQs

- "How accurate is the OCR now?"
- "What languages are supported?"
- "How do I correct an OCR error?"

Impact

Success metrics

- **KPI:** >95% OCR accuracy for English, >90% for Hindi and Mandarin on clear scanned documents.
- **KPI:** 25% reduction in average manual OCR review/correction time.

Next steps

- Monitor OCR performance in production.
- Gather user feedback on the OCR correction interface.
- Explore advanced OCR features (e.g., handwriting, form recognition) for future releases.