SMART-DOC-READER — PROMPT LOG

1. OCR PROCESSING PROMPT

Extract all readable text from the input image using Tesseract OCR engine.

Language: English (eng).

Return full text exactly as detected.

2. DOCUMENT CLASSIFICATION PROMPT

Classify the document into: statement / loan / unknown.

Loan keywords: loan, agreement, emi, borrower, lender, sanction, interest

Statement keywords: statement, deposits, withdrawals, balance, transactions

Return (docType, confidenceScore).

3. RULE-BASED EXTRACTION PROMPT

Extract fields using regex-based rules:

• accountNumber

• period

• avgBalance

• status

Confidence scoring:

+0.40 account

+0.30 period

+0.20 avgBalance

+0.10 status

If score $\geq 0.70 \rightarrow$ accept.

Else $\rightarrow$ LLM fallback.

4. LLM EXTRACTION PROMPT

Extract the following fields and return ONLY JSON:

[

accountNumber, period, avgBalance, status,

loanAmount, emiAmount, loanTenure,

borrowerName, lenderName, agreementDate

]

If missing → null. Return valid JSON only.

5. OFFLINE FALLBACK PROMPT

Use regex & heuristics if LLM not available:

• Account

• Period

• Loan Amount

• EMI

• Borrower

• Lender

Return normalized JSON.

6. API RESPONSE FORMAT PROMPT

Return JSON:

{

id,

docType,

docTypeConfidence,

file,

extracted,

confidence,

text_snippet

```
}
```