EMR Pipeline

Common pipeline stages based on the literature survey and adopted techniques:

1. **Document intake & triage (scan + classification)**
   - Scan documents at a consistent DPI (300 dpi typical).
   - Automatically classify document types (admission note, discharge summary, lab report) using ML models on image/OCR text
2. **Optical Character Recognition (OCR)**
   - Use a medical-aware OCR engine (or customize with zonal OCR and model fine-tuning for forms); expect printed text to be high accuracy, handwritten notes are still error-prone. Vendor and case studies recommend human-in-the-loop verification for critical fields.
3. **Preprocessing & OCR correction**
   - Normalize text (remove headers/footers, fix broken tokens, line-break handling), domain lexicons/medical dictionaries boost accuracy. Hsu et al. describe text cleaning steps before IE.
4. **Information Extraction (NLP / IE)**
   - Use a combination of rule-based extraction for standard structured fields and ML/NLP (NER, relation extraction) for narrative text. Evaluate with clinical NER metrics; modern LLMs can be promising but require careful prompting/fine-tuning and evaluation
5. **Normalization & coding**
   - Map extracted entities to controlled vocabularies (ICD-10, SNOMED CT, LOINC) and to FHIR resource fields (Patient, Observation, Condition). The FHIR mapping literature recommends visual/reusable transformation components for repeatable mappings.
6. **De-identification & privacy checks**
   - If data will be used for secondary purposes or shared, run clinical PHI detection and de-identification. Many papers and guides emphasize managing PHI before storage or research use.
7. **Integration / ingestion into EHR (FHIR/OpenEHR)**

- Transform structured outputs into FHIR resources or your target EHR schema. Use mapping tools and validate with schema and clinical stakeholders. FHIR is the recommended contemporary target for interoperability.

8. **Human-in-the-loop QA and continuous monitoring**

- Human review for sampled records, feedback loop to retrain models, monitor extraction accuracy and drift. This is emphasized across case studies.

Practical recommendations (based on the literature)

- Start with a **pilot** on a single document type (e.g., outpatient notes or lab reports). Papers that report success typically start small and expand.
- Use **zonal OCR** + dictionaries for forms; switch to advanced ML/NLP (NER or LLMs) for free-text extraction.
- Design mappings to **FHIR** from day one — even if you store data in a proprietary DB, mapping to FHIR later is much harder.
- Plan **human QC** for critical fields (drug names, allergies, diagnoses) — acceptable automation in literature always includes a human validation step.