Extract with Version 1	
Input	PDF of AE form
Textract Output	• Blocks (PAGE, LINE, WORD, KEY_VALUE_SET)
	• Bounding boxes, text content
Azure OpenAI Output	• Extracted fields (patient, drug, AE, dates)
	• Narrative sections
This is exactly what V1 does today	

Build Structural Representation	
From Textract + AOAI outputs, extract:	• Page count: 4 • Key labels & positions: - "Patient ID" at (0.1, 0.05) - "Adverse Event" at (0.1, 0.3) - "Drug Name" at (0.5, 0.2) • Section structure: - AE Summary (page 1) - Drug Info (page 2-3) - Reporter (page 4) • Tables detected: 2 - AE table (rows: 5, cols: 4) - Concomitant meds (rows: 8, cols: 3)

Serialize to canonical text:	"FormType=AE; Pages=4;
	Sections=[AE_Summary, Drug_Info, Reporter];
	KeyLabels=[Patient_ID@0.1-0.05,
	Adverse_Event@0.1-0.3, ...];
	Tables=[AE_Table(5x4), ConMeds(8x3)]"

Generate Embedding	
Take structural representation → OpenAI API	
Model:	text-embedding-3-small
Input:	Canonical text from Step 2
Output:	768-dimensional vector
Example: [0.023, -0.154, 0.891, 0.442, ...]	
This vector captures LAYOUT PATTERNS, not just text content	

KNN Search in Registry	

Query PostgreSQL with pgvector:	SELECT layout_id, layout_name,
	(1 - (embedding <=> $query_vector))
	AS similarity
	FROM layouts
	WHERE doc_type = 'AE_FORM'
	AND status = 'active'
	ORDER BY embedding <=> $query_vector
	LIMIT 5;

Results:	1. FDA_3500A_v2 (similarity: 0.94) ← WINNER 2. FDA_3500A_v1 (similarity: 0.89) 3. MedWatch_Custom (similarity: 0.72) 4. CIOMS_I (similarity: 0.65) 5. Custom_Form_X (similarity: 0.58)

Decision Logic	
IF top_similarity >= 0.75:	• Label: "Known Layout" • Action: Load policy for FDA_3500A_v2 • Policy includes: - Field mapping rules - Validation rules (date logic, drug checks) - Output format • Route: Automatic processing
ELSE (top_similarity < 0.75):	• Label: "Unknown/New Layout" • Action: Extract with generic policy • Route: HITL Review Queue • Reason: "Low confidence match (0.58)"