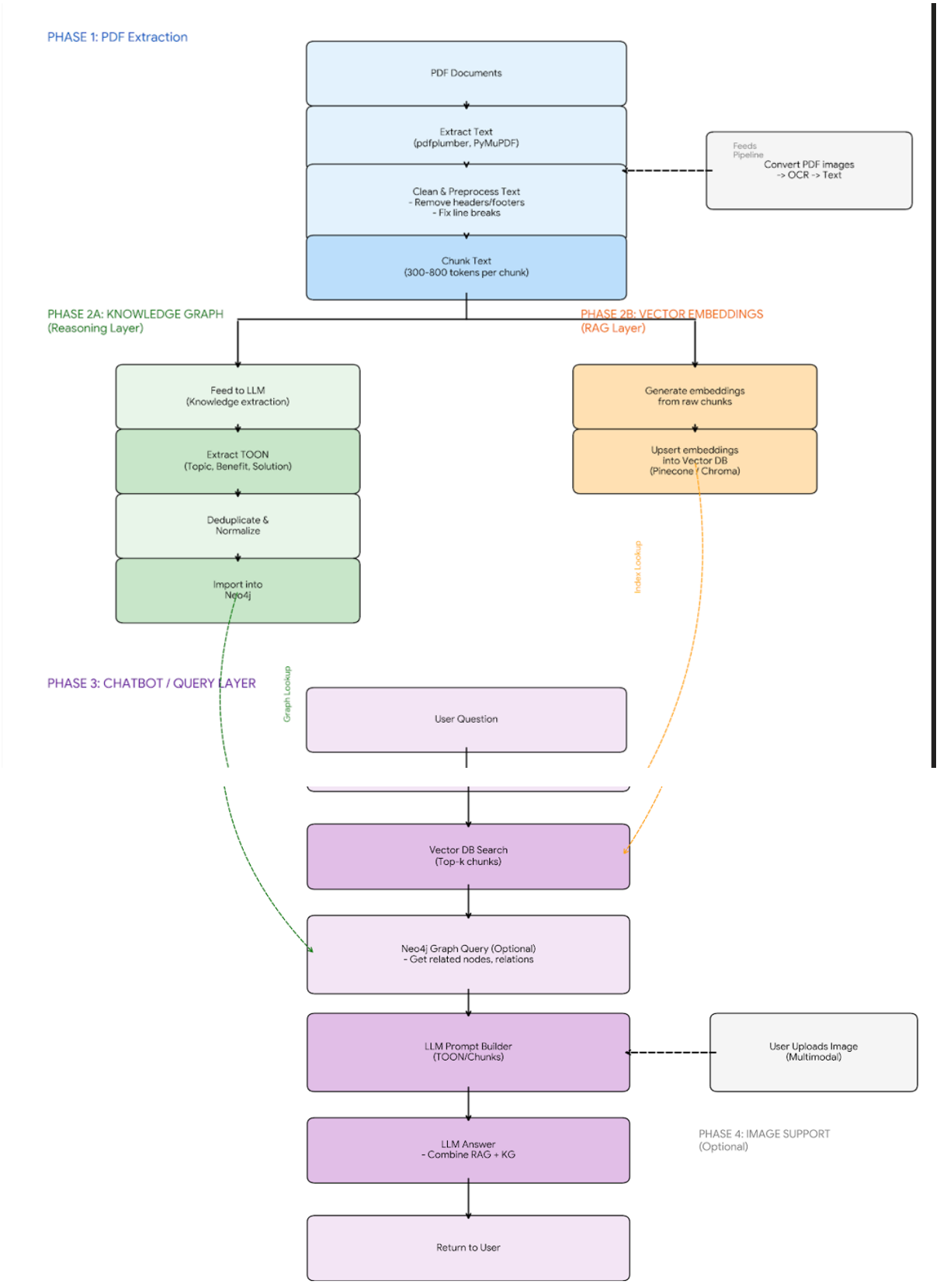
Name: Priyanshu Mishra



PHASE 1: PDF Extraction

Step	Purpose
PDF Documents	Source files
Extract Text	Get raw text from PDFs
Clean & Preprocess	Remove headers/footers, fix line breaks, normalize spacing
Chunk Text	Split into meaningful parts (300-800 tokens)

PHASE 2A: KNOWLEDGE GRAPH (Reasoning Layer)

Step	Purpose
Feed to LLM	Extract entities, topics, relations from chunk
Extract TOON	Structured output for KG
Deduplicate & Normalize	Merge similar concepts & standardize names
Import into Neo4j	Build graph DB

PHASE 2B: VECTOR EMBEDDINGS (RAG Layer)

Step	Purpose
Generate embeddings	Convert text chunks to vectors
Upsert embeddings	Store vectors for semantic search

PHASE 3: CHATBOT / QUERY LAYER

Step	Purpose
Vector DB Search	Retrieve top-k relevant chunks
Neo4j Graph Query (optional)	Retrieve structured relations for reasoning
LLM Prompt Builder	Combine RAG + KG info into prompt
LLM Answer	Generate final answer
Return to User	Serve chatbot response

PHASE 4: IMAGE SUPPORT (Optional)

Step	Purpose
Convert PDF images → OCR	Extract text from diagrams/images
Create text chunks → embed → RAG	Include image info in vector DB
User uploads image → LLM multimodal	Interpret uploaded images