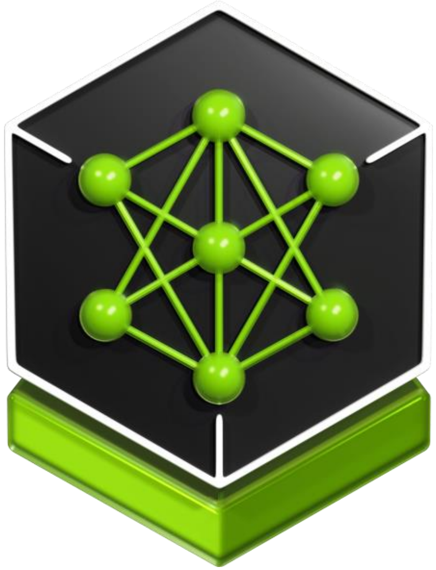
Next Steps

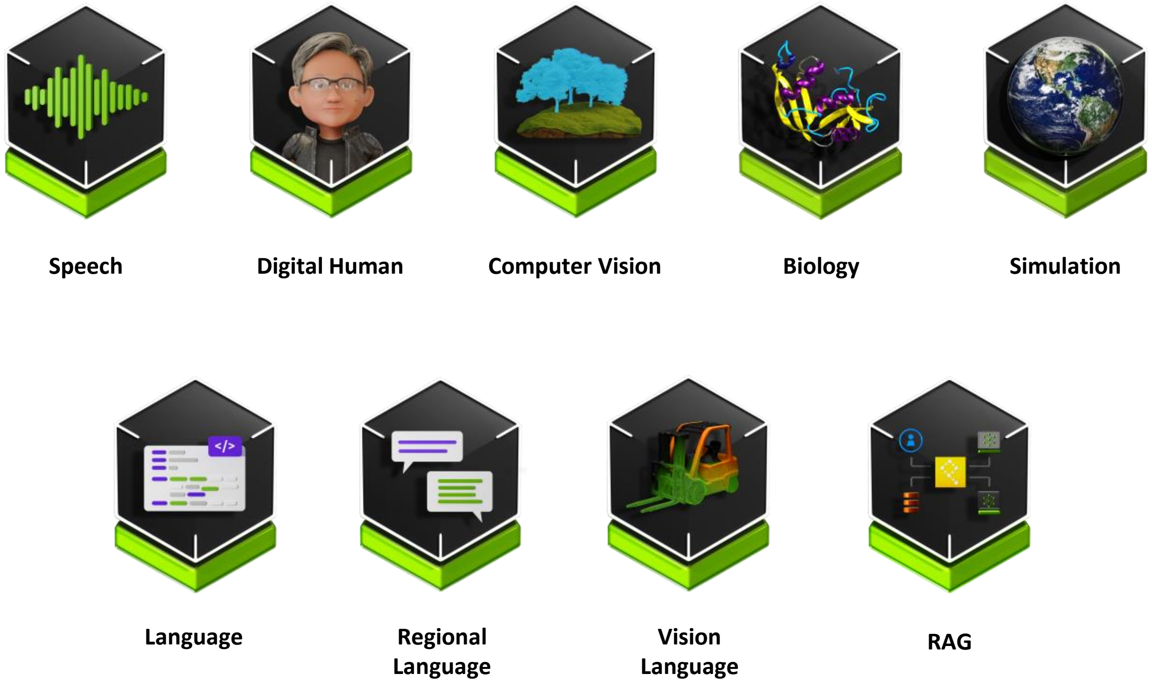
NVIDIA NIM: Optimized AI Models Run Up to 5X Faster

Community Models – Partner Models – NVIDIA Models



NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models
Packaged and Optimized to Run Across
CUDA Installed Base



Speech

Digital Human

Computer Vision

Biology

Simulation

Language

Regional Language

Vision Language

RAG

Next Steps

NVIDIA NIM Agent Blueprints

NIM Agent Blueprints

New blueprints releasing every month

1. Experience at build.nvidia.com

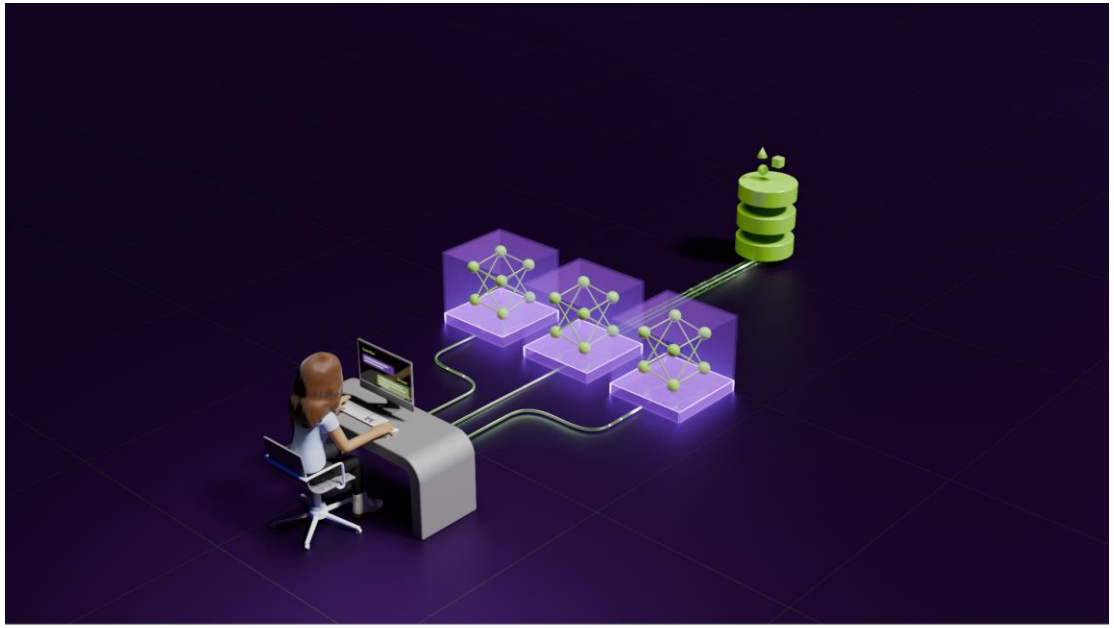
2. Download and run
on your own infrastructure

3. Contact an NVIDIA partner for assistance with customizing for your own business

LLM Tool Calling

Ability of LLMs to interact with external tools, APIs, or functions to perform tasks beyond text generation

- Equip NIM applications with agentic tools to enhance larger automation systems
- Enables autonomous agents and other AI applications to fetch real-time data, perform actions, and interact with external systems
  - Bridge the gap to new, real-world use cases that significantly enhance productivity and the user experience
- Steps:
  1. Defining tools
  2. Prompting the LLM
  3. Generating tool calls
  4. Executing tools
  5. Incorporating results into an LLM response



https://developer.nvidia.com/blog/spotlight-xpander-ai-equips-nvidia-nim-applications-with-agentic-tools/

Details are found here

NVIDIA NIM Agent Blueprints

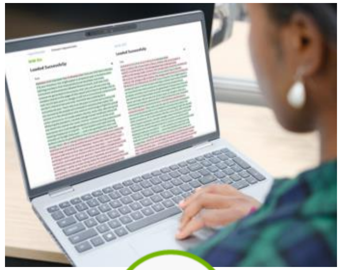
Reference AI Applications That Power Enterprises With Their Own AI Flywheel

NVIDIA NIM Agent Blueprints

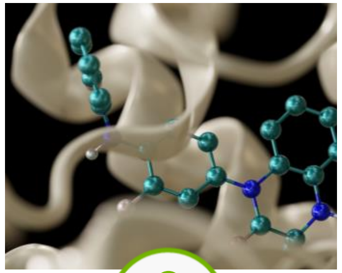
Available on build.nvidia.com

Digital Humans for Customer Service

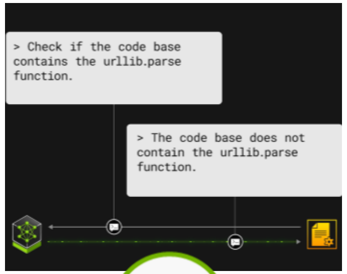
Multimodal PDF Data Extraction for Enterprise RAG



Generative Virtual Screening for Drug Discovery



Vulnerability Analysis for Container Security



```
> Check if the code base
contains the urllib.parse
function.
```

```
> The code base does not
contain the urllib.parse
function.
```

AI Virtual Assistants for Customer Service

Reference Application

Sample Data

Reference Code

Architecture

Customization Tools

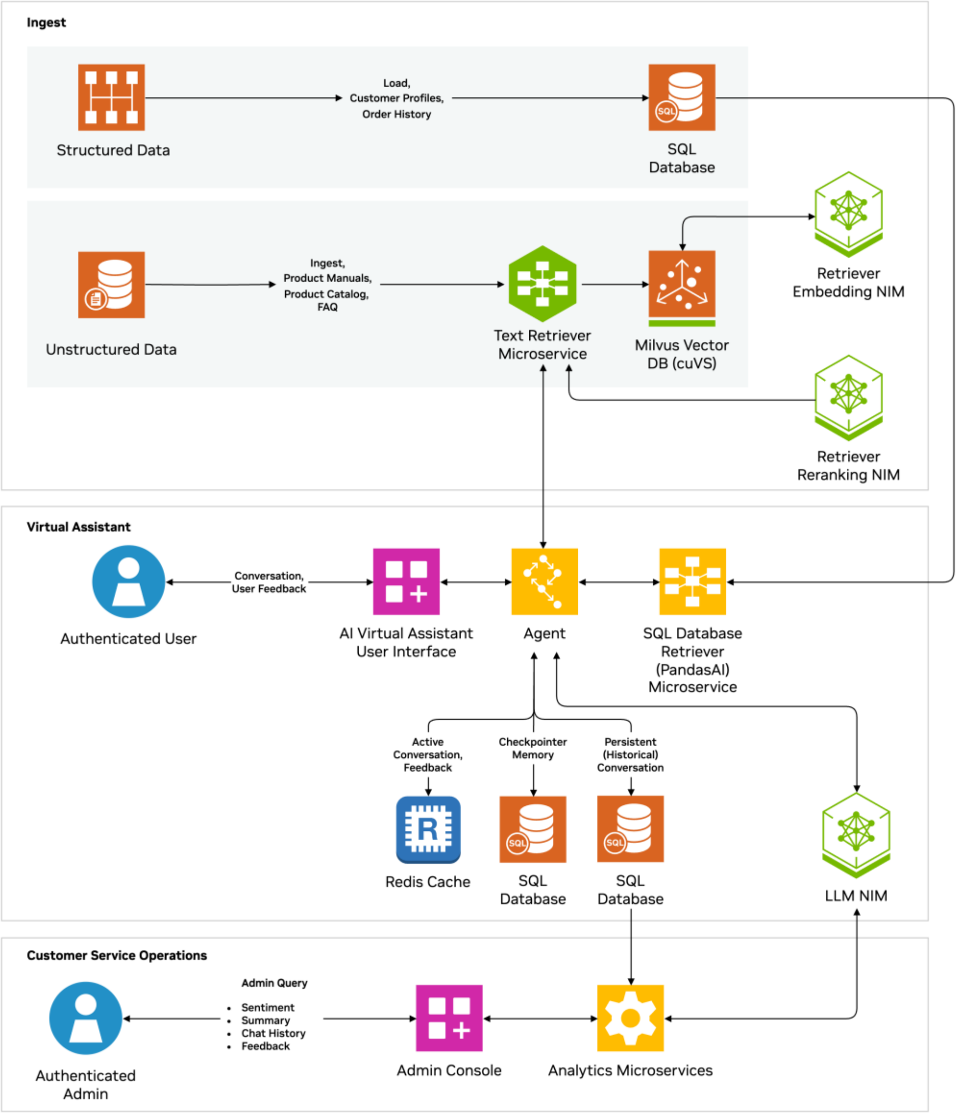
Orchestration Tools

AI Virtual Assistant for Customer Service

AI Virtual Assistant to reduce handling time, boost customer satisfaction

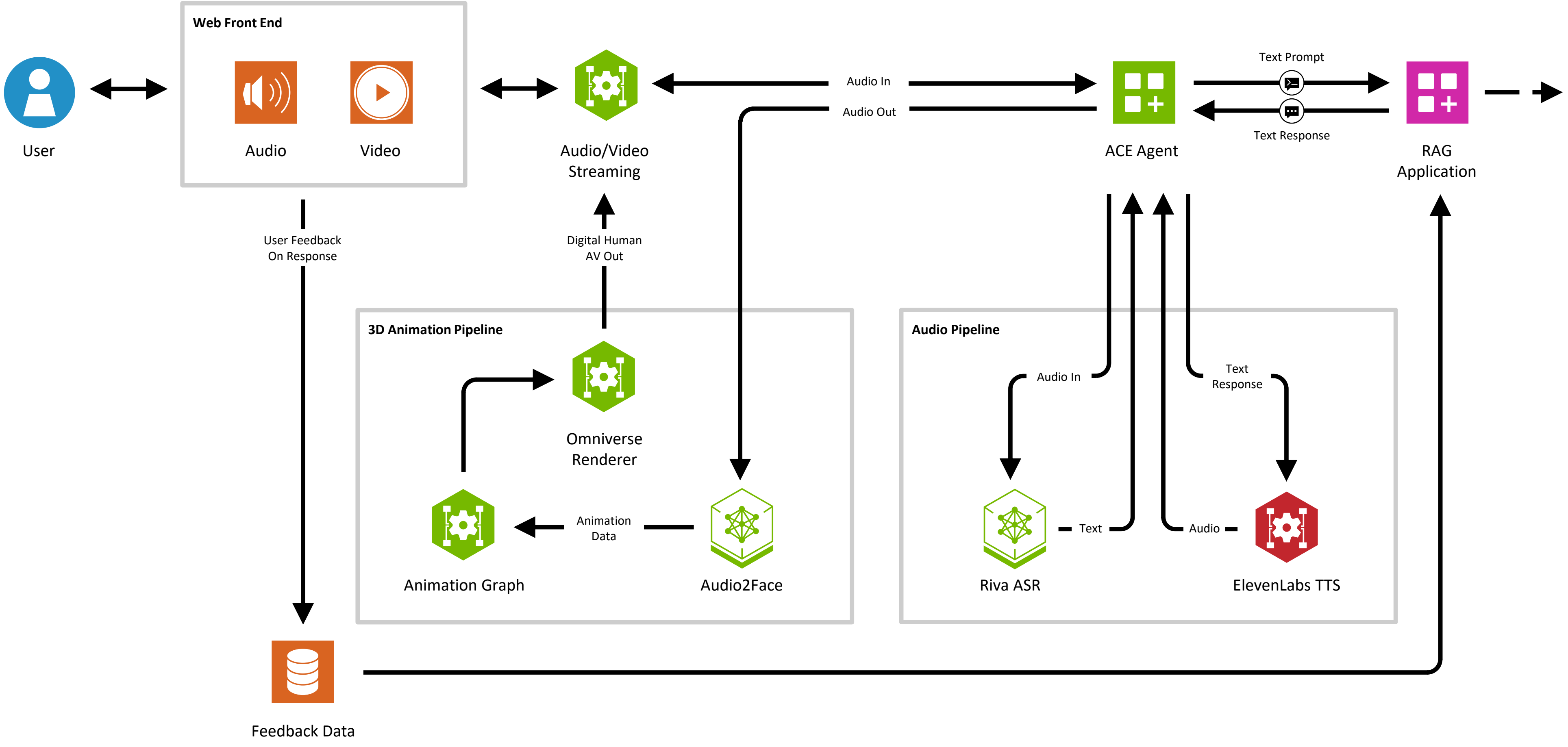
Benefits

- Personalized Responses: Handles structured and unstructured customer queries (e.g., order details, spending history).
- Multi-Turn Dialogue: Offers context-aware, seamless interactions across multiple questions.
- Custom Conversation Style: Adapts text responses to reflect corporate branding and tone.
- Sentiment Analysis: Analyzes real-time customer interactions to gauge sentiment and adjust responses.
- Multi-Session Support: Allows for multiple user sessions with conversation history and summaries.
- Data Privacy: Integrates with on-premises or cloud-hosted knowledge bases to protect sensitive data.



Digital Humans for Customer Service

$125B market for digital human economy by 2035



Benefits

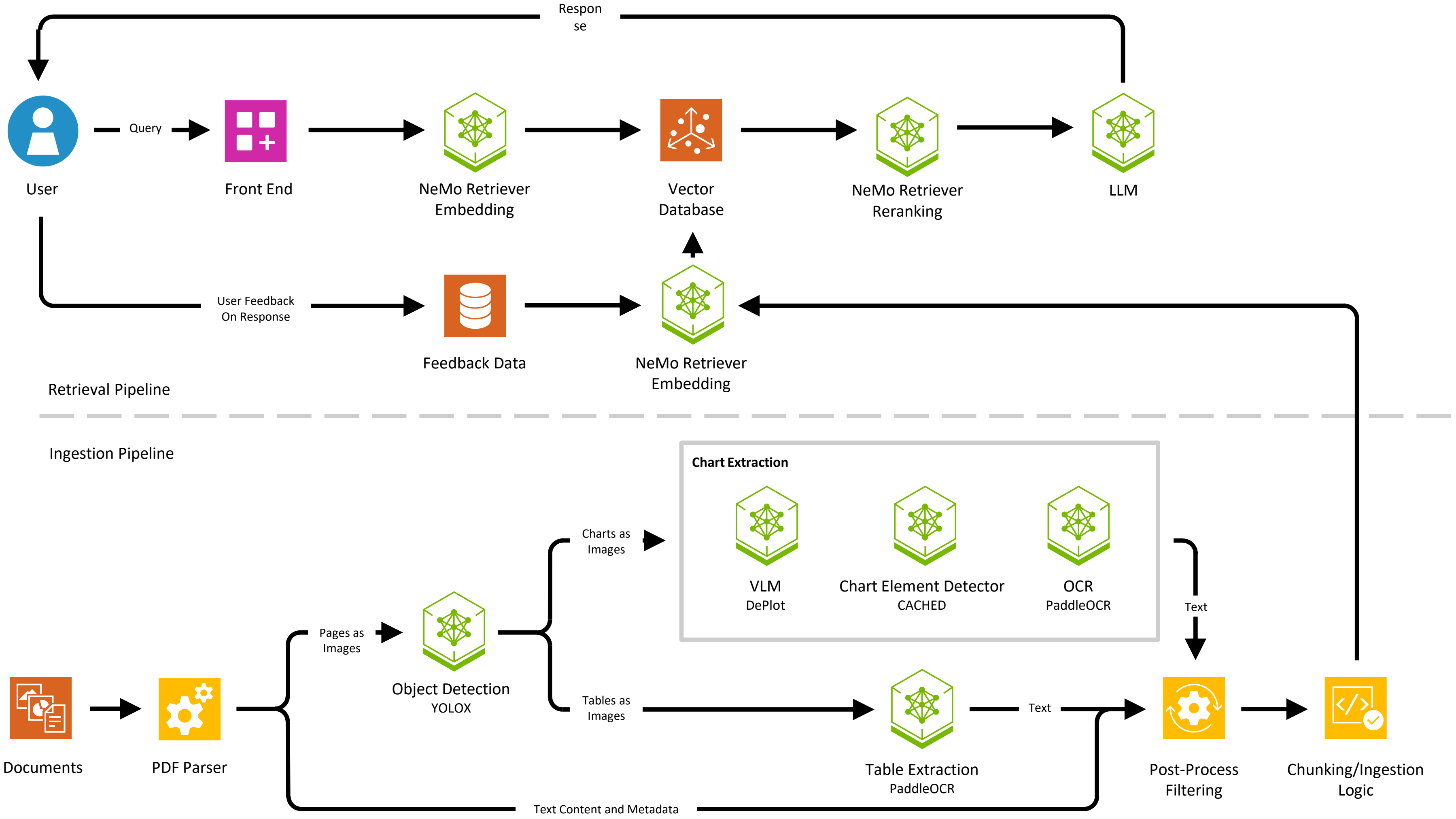
- Increases engagement and satisfaction for user-facing applications
- Creates a lifelike 3D digital human with accurate skin, hair, animation, and speech
- Enables natural conversations with enterprise applications and data

Multimodal PDF Data Extraction for Enterprise RAG

Unlocks Knowledge from trillions of PDFs

Benefits

- Unlocks the next level of indexable enterprise data from text to images and charts
- High-accuracy extraction and responses
- Enterprise-scale PDF ingestion



Respon
se
User
Query
Front End
NeMo Retriever Embedding
Vector Database
NeMo Retriever Reranking
LLM
User Feedback On Response
Feedback Data
NeMo Retriever Embedding
Retrieval Pipeline
Ingestion Pipeline
Chart Extraction
VLM
DePlot
Chart Element Detector
CACHED
OCR
PaddleOCR
Charts as Images
Text
Pages as Images
Object Detection
YOLOX
Tables as Images
Table Extraction
PaddleOCR
Text
Documents
PDF Parser
Post-Process Filtering
Chunking/Ingestion Logic
Text Content and Metadata