1. Chat files -

Data Source	Format	Chunking Method	Rationale	Strategy Name
Chats	.txt	• Fixed-size chunking with overlap splits where each chunk partially repeats part of the previous one to preserve context. • Participant names and their numbers are extracted too.	Context is not lost at chunk boundaries, and conversational flow is maintained.	Custom (General/Naive chunking strategy of ragflow does not allow chunk overlap)
Chats	.pdf	• Extract chat text from the PDF and structure it into messages. • Split the text into fixed-size chunks with partial overlap to keep context. • Identify unique participant names and count them.	Context is not lost at chunk boundaries, and conversational flow is maintained.	Custom
Chats	.json	• Dividing each message in json format • Grouping 5-10 messages together to form a chunk	Easy to process for the LLM and chunk size is consistent.	Custom
Chats	.xml	• Split the chat XML data into chunks by grouping a fixed number of<Message> elements (5-10 messages) together.	Maintains XML structure, easy for LLM to process and chunk size is consistent.	Custom

2. Bank Statements -

Data Source	Format	Chunking Method	Rationale	Strategy Name
Bank Statements	pdf	• Use extract_tables from PyMuPDF to extract structured transaction tables from bank statement PDFs. • Group extracted transactions into chunks of 5–10 rows, with some transactions overlapping between adjacent chunks to maintain context. • Attach the column headers to each transaction value in every chunk to ensure clarity and consistency. • Treat non structured data as a separate chunk	Preserves transactional context, and is scalable and robust.	Custom
Bank Statements	xlsx	• Group transactions into chunks of 5–10 rows each. • Include an overlap of a few transactions • Attach the column headers to each transaction value within every chunk for clarity.	Preserves transactional context and is scalable and robust.	Custom

Bank Statements	xls	• Group transactions into chunks of 5–10 rows each. • Include an overlap of a few transactions • Attach the column headers to each transaction value within every chunk for clarity • Treat non structured data as a separate chunk	Preserves transactional context and is scalable and robust.	Custom
Bank Statements	csv	• Group transactions into chunks of 5–10 rows each. • Include an overlap of a few transactions • Attach the column headers to each transaction value within every chunk for clarity • Treat non structured data as a separate chunk	Preserves transactional context and is scalable and robust.	Custom
Bank Statements	html	• Group transactions into chunks of 5–10 rows each. • Include an overlap of a few transactions • Attach the column headers to each transaction value within every chunk for clarity • Treat unstructured data separate chunk	Preserves transactional context and is scalable and robust.	Custom

3. Video

Data Source	Format	Chunking Method	Rationale	Strategy Name
Video	.video/*	● First the video can be transcribed using ASR and then ● semantic chunking can be done, i.e. similar sentences can be chunked together (sentence transformers) ● can add time stamp based overlap	Preserves flow and context, retrieval is good.	Custom

4. Image

Data Source	Format	Chunking Method	Rationale	Strategy Name
Image	.image/*	● For retrieval, slide fixed size window over image, convert to embeddings and store them ● For Ocr, run ocr first and store similar to pdf	Preserves flow and context, retrieval is good.	Custom

5. Audio

Data Source	Format	Chunking Method	Rationale	Strategy Name
Audio	.audio/*	● First the video can be transcribed using ASR and then ● semantic chunking can be done, i.e. similar sentences can be chunked together (sentence transformers) ● can add time stamp based overlap	Preserves flow and context, retrieval is good.	Custom

6. Others

Data Source	Format	Chunking Method	Rationale	Strategy Name
Others	pdf	● PDF can be parsed into structured data using NER (used for resumes) ● Using the lowest section titles as the basic unit for chunking documents. Therefore, figures and tables in the same section will not be separated, which may result in larger chunk sizes. (for manuals) ● For research papers, it	● NER extracts key fields as meaningful chunks for precise data capture. ● Lowest section titles group related content, preserving instructional integrity.	● Resume ● Manual ● Paper ● Laws ● Present ation

		can be chunked using sections (eg, abstract, intro) ● Legal documents are chunked by detecting structured text features like hierarchical headings (e.g., "ARTICLE") as split points, preserving document structure. ● In case of presentations, every slide is treated as a separate chunk.	● Results in good info retrieval ● Hierarchical headings keep legal provisions intact in chunks. ● Each slide stays a clear, self-contain ed content unit.	
Others	xlsx	● For Q&A documents, each qa pair is treated as one chunk ● For table data, each row is considered as a separate chunk	● Context between each pair is preserved	● Q&A ● Table
Others	xls	● For Q&A documents, each qa pair is treated as one chunk ● For table data, each row is considered as a separate chunk	● Context between each pair is preserved	● Q&A ● Table
Others	csv	● For Q&A documents, each qa pair is treated as one chunk ● For table data, each row is considered as a separate chunk	● Context between each pair is preserved	● Q&A ● Table

Others	rtf	• Split the text based on its hierarchical structure (headings and sections), preserving logical divisions and optionally using overlap to maintain context.	• Preserves the logical flow of the document.	Custom
Others	txt	Same as PDF, only extraction will be different		
Others	docx	Same as PDF, only extraction will be different		
Others	json	• Dividing each text content present in json format • Grouping 5-10 messages together to form a chunk	Easy to process for the LLM and chunk size is consistent.	Custom
Others	xml	• Split the chat XML data into chunks by grouping a fixed number of<Text> elements (5-10 messages) together.	Maintains XML structure, easy for LLM to process and chunk size is consistent.	Custom
Others	html	• Can use HTMLSemanticPreservingSplitter from langchain which splits html content correctly, while keeping semantic meaning intact	Context is maintained	Custom

7. Knowledge Base

Data Source	Format	Chunking Method	Rationale	Strategy Name
Knowledge Base	.pdf	● Fixed-size chunking with overlap splits where each chunk partially repeats part of the previous one to preserve context	Context is not lost at chunk boundaries, and info flow is maintained.	Custom
Knowledge Base	.txt	● Fixed-size chunking with overlap splits where each chunk partially repeats part of the previous one to preserve context	Context is not lost at chunk boundaries, and info flow is maintained.	Custom

8. Email

Data Source	Format	Chunking Method	Rationale	Strategy Name
Email	.pst	● Extract each email from the .pst file (pypff) ● Group convo thread or subject ● Chunk by N consecutive emails in that thread ● Retain metadata (sender name etc) ● Overlap of emails	Conversational flow is maintained and overlap ensures no context loss.	Custom

Email	.eml	• Extract each email from the .pst file (mailbox) • Group convo thread or subject • Chunk by N consecutive emails in that thread • Retain metadata (sender name etc) • Overlap of emails	Conversational flow is maintained and overlap ensures no context loss.	Custom
Email	.olm	• Extract each email from the .pst file (extract_msg) • Group convo thread or subject • Chunk by N consecutive emails in that thread • Retain metadata (sender name etc) • Overlap of emails	Conversational flow is maintained and overlap ensures no context loss.	Custom
Email	.mbox	• Extract each email from the .pst file (olmreader) • Group convo thread or subject • Chunk by N consecutive emails in that thread • Retain metadata (sender name etc) • Overlap of emails	Conversational flow is maintained and overlap ensures no context loss.	Custom

9. Packet Capture Format

Data Source	Format	Chunking Method	Rationale	Strategy Name
Packet Capture Format	.pcap	• Group packets into chunks of 5–10 rows each, of every 5-10 seconds. • Include an overlap of a few packets • Attach the column headers to each for clarity.	Preserves context and is scalable and robust.	Custom