flowchart TB

A["Cleaned DataFrame"]

%% User Choice (no shaded box)

B{"User Enabled: Choose Chunking Method"}

%% Chunking Methods (aligned horizontally)

C1["Apply Fixed Chunking"]

C2["Apply Recursive Chunking"]

C3["Apply Semantic Chunking"]

C4["Apply Document Chunking"]

C5["Apply Agentic Chunking"]

%% Execution (aligned horizontally)

D1["Create Fixed Chunks"]

D2["Create Recursive Chunks"]

D3["Create Semantic Chunks"]

D4["Create Document Chunks"]

D5["Create Agentic Chunks"]

E["Final List of Text Chunks"]

F["Attach Metadata"]

G["📦 Send to Embedding Model"]

%% Main Flow

A --> B

B --> C1 --> D1 --> E

B --> C2 --> D2 --> E

B --> C3 --> D3 --> E

B --> C4 --> D4 --> E

B --> C5 --> D5 --> E

E --> F --> G

%% Force horizontal alignment

{rank=same; C1; C2; C3; C4; C5}

{rank=same; D1; D2; D3; D4; D5}

%% Styling

classDef dataFill fill:#e1f5fe,stroke:#01579b,stroke-width:2px,color:#000

classDef strategyFill fill:#fff3e0,stroke:#f57c00,stroke-width:2px,color:#000

classDef processFill fill:#e8f5e9,stroke:#388e3c,stroke-width:2px,color:#000

classDef outputFill fill:#f5f5f5,stroke:#616161,stroke-width:2px,color:#000

classDef decisionFill fill:#ffecb3,stroke:#ffa000,stroke-width:2px,color:#000

A:::dataFill

B:::decisionFill

C1:::strategyFill

C2:::strategyFill

C3:::strategyFill

C4:::strategyFill

C5:::strategyFill

D1:::processFill

D2:::processFill

D3:::processFill

D4:::processFill

D5:::processFill

E:::outputFill

F:::outputFill

G:::dataFill

%% Force straight orthogonal arrows

linkStyle default interpolate linear

Chunking

flowchart TD

A["📁 Raw CSV File"]

%% Structural Validation

subgraph B["Structural Validation"]

direction TB

B1["Detect Encoding"]

B2["Identify Delimiter"]

B3["Validate Headers"]

end

C["Load into DataFrame"]

%% Column Processing

subgraph D["Column Processing"]

direction TB

D1["Text Columns"]

D2["Numeric Columns"]

D3["Unique Values"]

end

%% Text Cleaning

subgraph E["Text-Specific Cleaning"]

direction TB

E1["Convert to String - Preserve NaN"]

E2["Lowercase & Strip"]

E3["Remove Extra Whitespace"]

E4["Remove HTML Tags"]

end

%% Conditional Metadata Enrichment

subgraph L["Optional Metadata Enrichment"]

direction TB

M["Named Entity Recognition"]

N["Keyword Extraction"]

O["Sentiment Analysis"]

end

H["🔄 Final Clean DataFrame"]

I{"Content Analysis?"}

F["Preserve Numeric as Metadata"]

G["Preserve Unique Values as Metadata"]

P["Add to Metadata"]

Q["➡️ Proceed to Chunking"]

K["Skip Sentiment Analysis"]

%% Main Flow

A --> B --> C --> D

D1 --> E --> H

D2 --> F --> P

D3 --> G --> P

H --> I

I -->|"Subjective - reviews or feedback"| O

I -->|"Objective - logs or specs"| K

O --> P

L --> P

P --> Q

%% Styling

classDef structuralFill fill:#f3e5f5,stroke:#7b1fa2,stroke-width:2px,color:#000

classDef dataFlowFill fill:#e1f5fe,stroke:#01579b,stroke-width:2px,color:#000

classDef processFill fill:#e8f5e9,stroke:#388e3c,stroke-width:2px,color:#000

classDef optionalFill fill:#fff3e0,stroke:#f57c00,stroke-width:2px,color:#000

classDef outputFill fill:#f5f5f5,stroke:#616161,stroke-width:2px,color:#000

classDef decisionFill fill:#ffecb3,stroke:#ffa000,stroke-width:2px,color:#000

B1:::structuralFill

B2:::structuralFill

B3:::structuralFill

C:::dataFlowFill

D1:::processFill

D2:::processFill

D3:::processFill

E1:::processFill

E2:::processFill

E3:::processFill

E4:::processFill

F:::outputFill

G:::outputFill

H:::outputFill

I:::decisionFill

M:::optionalFill

N:::optionalFill

O:::optionalFill

P:::outputFill

Q:::outputFill

K:::outputFill

Preprocessing

flowchart TB

%% User

User([User]) --> App[Single Application]

%% Subgraph for UI Components

subgraph UI[UI Components]

direction TB

FU[File Uploader]

NC[New CSV]

PL2[Preprocessing Layer 2]

CTS[Chunking Type Selection]

SB[Search Bar]

FO[Filter Options]

RD[Results Display]

end

%% Subgraph for Core Logic

subgraph CL[Core Logic]

direction TB

PUC[Process Uploaded CSV]

CH[Chunking]

GE[Generate Embeddings]

QP[Query Processing]

end

%% Subgraph for Data Store

subgraph DS[Data Store]

direction TB

DB[(ChromaDB In-Memory)]

end

%% Connections (straight arrows only)

App --> FU --> NC --> PL2 --> CTS --> PUC

PUC --> CH --> GE --> DB

QP --> DB

DB --> QP

SB --> QP

FO --> QP

QP --> RD

%% Styles

classDef ui fill:#e3f2fd,stroke:#1565c0,stroke-width:2px,color:#000,rx:6px,ry:6px;

classDef core fill:#e8f5e9,stroke:#2e7d32,stroke-width:2px,color:#000,rx:6px,ry:6px;

classDef data fill:#fffde7,stroke:#f9a825,stroke-width:2px,color:#000,rx:6px,ry:6px;

classDef general fill:#f5f5f5,stroke:#424242,stroke-width:2px,color:#000,rx:6px,ry:6px;

%% Assign classes

class FU,NC,PL2,CTS,SB,FO,RD ui;

class PUC,CH,GE,QP core;

class DB data;

class User,App general;

Ui