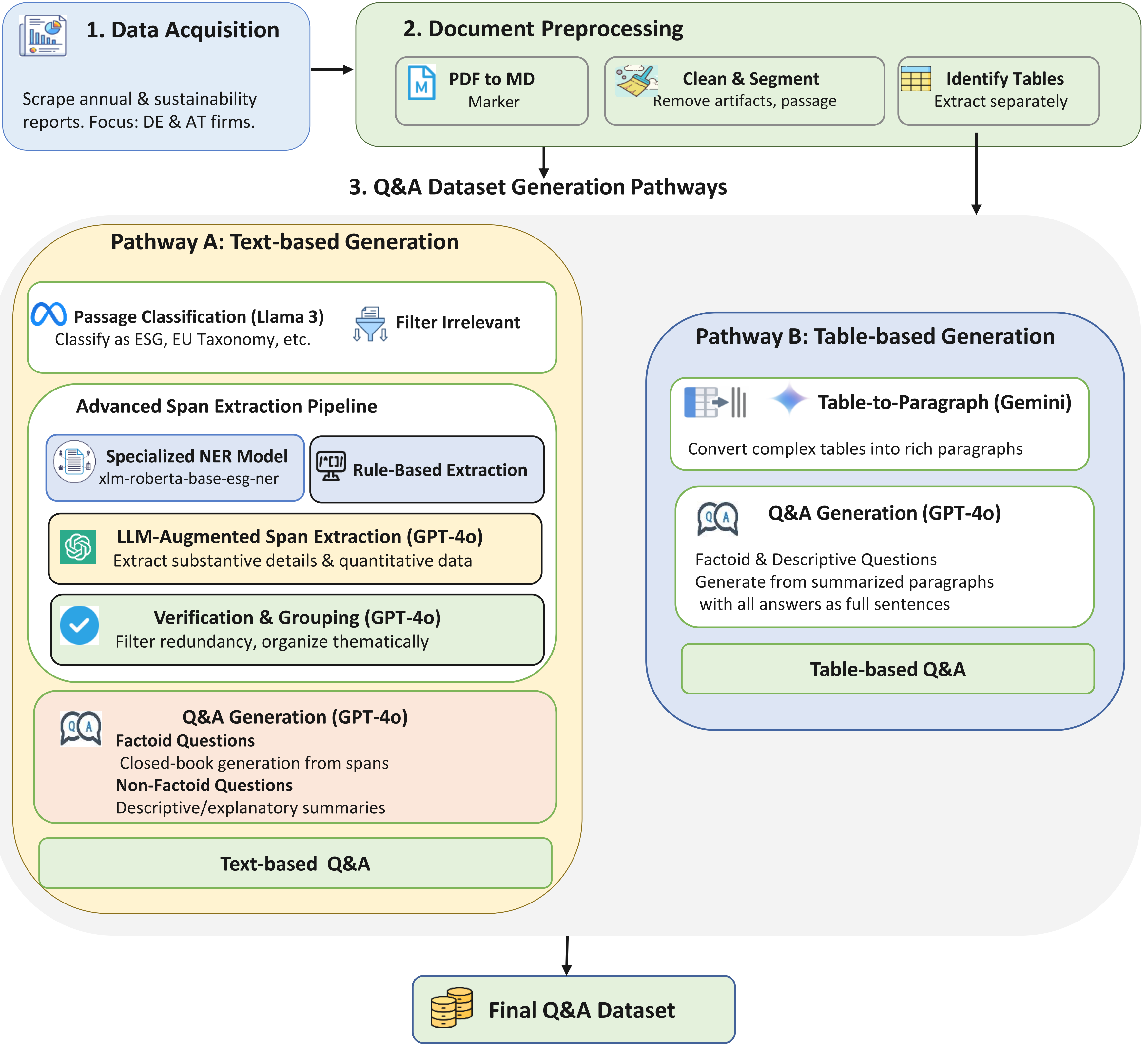

1. Data Acquisition

Scrape annual & sustainability reports. Focus: DE & AT firms.

2. Document Preprocessing

PDF to MD
Marker

Clean & Segment
Remove artifacts, passage

Identify Tables
Extract separately

3. Q&A Dataset Generation Pathways

Pathway A: Text-based Generation

Passage Classification (Llama 3)
Classify as ESG, EU Taxonomy, etc.

Filter Irrelevant

Advanced Span Extraction Pipeline

Specialized NER Model
xlm-roberta-base-esg-ner

Rule-Based Extraction

LLM-Augmented Span Extraction (GPT-4o)
Extract substantive details & quantitative data

Verification & Grouping (GPT-4o)
Filter redundancy, organize thematically

Q&A Generation (GPT-4o)
Factoid Questions
Closed-book generation from spans
Non-Factoid Questions
Descriptive/explanatory summaries

Text-based Q&A

Pathway B: Table-based Generation

Table-to-Paragraph (Gemini)
Convert complex tables into rich paragraphs

Q&A Generation (GPT-4o)
Factoid & Descriptive Questions
Generate from summarized paragraphs
with all answers as full sentences

Table-based Q&A

Final Q&A Dataset