Chunking Strategy Analysis Report

This report summarizes chunking strategy results for a structured PDF document and a conversational podcast transcript, focusing on boundary preservation, semantic coherence, and RAG suitability.

Chunking Strategy Recommendations

For PDF Documents

Recommended Strategy: Recursive Character Chunking

Reasoning:
- Preserves document structure by prioritizing paragraph and newline boundaries before splitting at sentence or word level
- Significantly reduces mid-word and mid-sentence splits compared to fixed-size chunking
- Maintains semantic coherence of sections, headings, and structured formatting
- Optimal chunk size: 1000 characters with 200 character overlap to balance structure preservation and manageable chunk count

For Podcast Transcripts

Recommended Strategy: Token-Based Chunking

Reasoning:
- Aligns directly with LLM context window limits, enabling more predictable embedding and retrieval behavior
- Produces fewer, larger context-rich chunks that make more efficient use of model context
- Recursive splitting provides limited improvement due to the continuous conversational structure of spoken language
- Optimal chunk size: 500 tokens with 50 token overlap to balance continuity and efficiency

Trade-offs Summary

Strategy	Pros	Cons	Best For
Fixed-Size	Simple and predictable implementation	Breaks sentences and structure	Uniform or baseline use
Recursive	Preserves structural boundaries	Slightly variable chunk sizes	Structured documents
Token-Based	Accurate for LLM context windows	Not structure-aware	RAG systems, transcripts
Semantic	Meaning-based splitting	Computationally expensive	Complex long-form text