**Justifying Hybrid Chunking Strategy**

*Implemented using semantic grouping + fixed token limits with transformers & tiktoken.*

* **What is the Hybrid Chunking Strategy?**

1. Semantic grouping of meaningful paragraphs.
2. Fixed token cap (512 tokens per chunk) to ensure

LLM compatibility and indexing efficiency.

* **Code Summary:**

1. Split text into meaningful paragraphs
2. Embed paragraphs using a Sentence Transformer model
3. Group semantically-related paragraphs without exceeding max token limit (512)

* **Performance Justification**

1. *Retrieval Efficiency*:
2. Token limit ensures all chunks are accepted by embedding models and Pinecone
3. Fixed-size grouping = consistent, fast vector retrieval
4. Smaller and coherent chunks avoid memory overload during search
5. Semantic Coherence:
6. Paragraphs grouped meaningfully using natural structure.
7. Semantics preserved better than fixed-size-only method.
8. Avoids breaking mid-topic or legal definitions.
9. Pinecone Compatibility:
   1. All chunks are below token threshold (e.g., 512), ensuring smooth vector indexing.
   2. Works with Pinecone's metadata filters (e.g., section tags) if added.