Report & Recommendations

* Summarisation that includes:

1. Comparison of chunking strategies.
2. Justification of the best approach.
3. Potential improvements and future considerations.
4. Comparison of chunking strategies.
5. **Fixed-size token-based chunking**

This method is fast and easy to implement, ensuring consistent chunk sizes and compatibility with token-limited models. However, it often breaks sentences or logical sections midway, reducing semantic coherence and retrieval relevance. While it's efficient, it risks context fragmentation and slight overlap redundancy.

1. **Sentence/paragraph-based chunking**

Sentence or paragraph-based chunking preserves natural language boundaries, improving relevance and readability compared to fixed-token splits. Though it sacrifices chunk size consistency and is slower with dense texts, it enhances contextual flow during retrieval.

1. **Section-based chunking using headers and metadata**

Section-based chunking aligns splits with document structure to reduce fragmentation and improve coherence, especially in legal PDFs. It enhances retrieval and supports metadata-rich queries but depends on accurate formatting and header detection, adding complexity.

1. **Semantic chunking**

Semantic chunking uses embeddings to group similar text, overcoming section-based limits when headers are missing. It boosts cohesion and relevance in unstructured content but is slower, less predictable in size, and trades control for deeper, context-aware responses.

1. **Late chunking**

Late chunking delays splitting until after retrieval, enabling dynamic, query-based chunks for better accuracy and less context loss. Though it increases latency and slows indexing, it's ideal for precise tasks like summarization or QA.

1. **Contextual Retrieval by Anthropic**

Contextual Retrieval by Anthropic selects relevant spans dynamically using the full query and conversation history, unlike late chunking which still slices post-retrieval. It boosts relevance and coherence but relies on advanced, resource-heavy models—shifting focus from chunking to smarter retrieval.

1. **Hybrid approaches**

The hybrid approach combines semantic awareness with fixed-token control, balancing coherence, consistency, and efficiency. It avoids high compute costs of methods like Anthropic’s, offering a practical, context-rich solution for complex, structured documents.

1. Justification of the best approach
2. Combines semantic understanding with token control for both meaning and model efficiency.
3. Works well even when headers or document structure are inconsistent.
4. Balances semantic coherence, chunk size consistency, and retrieval accuracy.
5. Keeps chunks within token limits, reducing context loss.
6. Suitable for real-world use, efficient, adaptable, and production-friendly.
7. Potential improvements and future considerations

The hybrid search approach can be improved by fine-tuning embeddings for domain-specific accuracy, using dynamic chunking based on semantic importance, adding fallback strategies for edge cases, and refining performance through real user feedback.