

Semantic Chunk Coherence

SCC

1 - 10

Measures whether a content unit is a clean, single-topic, self-contained chunk — one that an AI retrieval system can extract, process, and return without requiring surrounding context to interpret it.

Weight: 0.20 in RASA composite

WHAT IT MEASURES

SCC governs whether a content chunk embeds cleanly in vector space. Multi-topic chunks embed to blurred vectors, positioned ambiguously between topic centroids — competitive for none. Clean SCC ensures accurate retrieval and reliable synthesis regardless of RP score.

GOVERNS

- RAG Chunk Indexing
- Vector Embedding Accuracy
- Passage Extraction
- AI Answer Synthesis Quality

4 PRIMARY STRUCTURAL FACTORS THAT DRIVE SCC SCORES

F-01

Single-Topic Focus

Each chunk must address one clearly scoped concept, question, or claim. Multi-topic chunks force AI systems to make ambiguous embedding judgements, reducing retrieval precision.

F-02

Logical Internal Flow

Content should follow a coherent sequence — definition, elaboration, evidence, conclusion — rather than jumping between sub-points or switching register mid-passage.

F-03

Self-Containment

A high-SCC chunk is interpretable without reading surrounding content. No pronoun references to prior context, no undefined key terms, no conclusions from elsewhere.

F-04

Absence of Topic Drift

Drift occurs when a chunk migrates from one topic to another via loose transitions ("It is worth noting that..."). This splits the embedding and reduces precision for both topics.

COMMON FAILURE MODES

F1

Fragmented Info Structures

Bullet-heavy or heading-dense layouts interrupt semantic flow. Each fragment may be accurate but lacks internal logic for coherent chunk retrieval.

F2

Mixed-Intent Passages

Simultaneously defining a term, making a sales argument, and citing a case study in one paragraph — conflicting intents prevent clean embedding.

F3

Context-Dependent Openings

Chunks opening with "As discussed above..." or "Building on this..." carry broken references when extracted by a RAG pipeline, degrading synthesis quality.

SCC SCORE REFERENCE SCALE

RASA COMPOSITE FORMULA

$$\text{RP} \times 0.25 + \text{SCC} \times 0.20 \\ + \text{ECS} \times 0.20 + \text{SCI} \times 0.20 \\ + \text{CGP} \times 0.15$$

PUBLISH ≥ 8.0 composite
REJECT < 6.0 on SCC alone

1 – Incoherent 3 – Weak 5 – Moderate 7 – Strong 9-10 – Exceptional

9-10

EXCEPTIONAL

"The SCC metric evaluates single-topic focus: each chunk addresses one concept, enabling precise vector embedding and accurate AI retrieval."

7-8

STRONG

"Semantic chunking improves retrieval accuracy. Metadata tags can also help categorise content for downstream use."

5-6

MODERATE

"Good content needs to be well-written and relevant. AI-readiness and SEO best practices should also be applied."

3-4

WEAK

"Building on the previous section, this relates to what was discussed earlier, which also connects to the broader topic."

1-2

INCOHERENT

"This means that, as noted above, the approach — which varies — may or may not apply depending on context."

SCC IMPROVEMENT CHECKLIST

- Write each section to address exactly one scoped question or concept
- Open every chunk with a self-contained statement — no prior context references
- Remove transitions that introduce new topics mid-chunk
- Split any passage covering more than one distinct topic into separate chunks
- Ensure the last sentence does not depend on following content to complete its meaning
- Test self-containment by reading the chunk in isolation before publishing

DD1

Retrieval Probability

DD2 · ACTIVE

Semantic Chunk Coherence

DD3

Entity Clarity Score

DD4

Synthesis Compatibility Index

DD5

Citation & Grounding Potential