

Adaptive Chunking: OPTIMIZING CHUNKING-METHOD SELECTION FOR RAG

Moura Júnior, Paulo Roberto · Jean Lelong · Annabelle Blangero

Ekimetrics.

LREC 2026

github.com/ekimetrics/adaptive-chunking

Motivation

The effectiveness of RAG is highly dependent on how documents are **chunked**: segmented into smaller units for indexing and retrieval. Yet, commonly used “one-size-fits-all” approaches break down across the structural and semantic variety of real documents.

Despite its central role, chunking lacks a **dedicated evaluation framework**, making it difficult to assess and compare strategies independently of downstream performance.

We introduce **Adaptive Chunking**: a framework that selects the most suitable chunking strategy for each document, guided by five intrinsic metrics.

New Chunking Methods

1. LLM-Regex Splitter

- LLM analyzes structure → generates a **document-specific regex**
- Deterministic `re.split()`: fast and reproducible
- Best for structured texts (legal, numbered sections)

2. Split-then-Merge Recursive Splitter

- **Pass 1**: hierarchical split (titles → sections → sentences → chars)
- **Pass 2**: greedy merge with overlap
- Variants at max 1,100 and max 600 tokens

Post-processing

- Re-split chunks >1,100 tokens; merge chunks <100 tokens

Evaluation Corpus

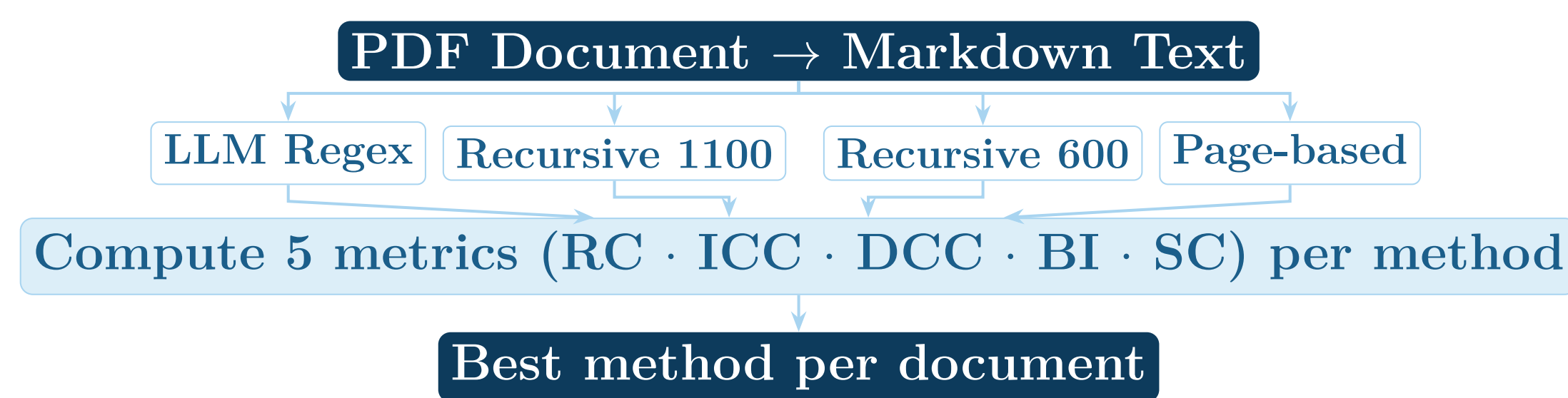
33 PDFs from the Ekimetrics CLAIR project, **~1.18M tokens**, parsed with Azure Document Intelligence.

Domain	Docs	Tok/doc	Pg/doc
Technical	9	5,257	12
Legal	16	30,895	46
Social Sci.	8	79,862	114

Two-track evaluation (*guards against self-validation by our own metrics*)

- **Intrinsic**: our 5 metrics, on chunks alone.
- **Extrinsic**: hybrid RAG (sparse + dense + reranker, $k=10$) with gpt-4.1 generator, LLM-as-judge on 99 expert QA pairs.

Adaptive Chunking Pipeline



Per-document selection: no single method is universally optimal. Runs at indexing time, with no query-time overhead.

5 Intrinsic Chunking Metrics

Note: simplified examples for illustration purposes.

RC — References Completeness

Fraction of entity-pronoun pairs preserved within the same chunk. Uses Maverick coreference resolution.

× **BROKEN**
Chunk 1: "... **The Hamburg Commissioner** published guidelines on LLM data processing. ..."
Chunk 2: "... **They** mandate encryption for all personal data transfers. ..."

INTACT
"... **The Hamburg Commissioner** published guidelines on LLM data processing. **They** mandate encryption for all personal data transfers. ..."

ICC — Intrachunk Cohesion

Mean cosine similarity between sentence embeddings and full chunk embedding (Jina v3). Higher = single topic per chunk. *Favours smaller, focused chunks.*

× **LOW COHESION**
"Art. 5 defines data minimization. [...] Table 3: GDP growth rates for Q2 2024 across EU member states."

HIGH COHESION
"Art. 5 defines data minimization. Personal data must be adequate, relevant, and limited to what is necessary."

DCC — Document Contextual Coherence

Mean cosine similarity between each chunk and its 3,000-token sliding context window. Ensures chunks are understandable in context. *Favours larger chunks; balances ICC.*

× **LOW COHERENCE**
Chunk about "lawfulness of processing" surrounded by context about financial regulations → low alignment with window

HIGH COHERENCE
Chunk about "lawfulness of processing" surrounded by GDPR articles on data protection → high alignment with window

BI — Block Integrity

Proportion of structural units (paragraphs, tables, title-body pairs) that remain unbroken across chunk boundaries. Block spans are provided by the parsing step.

× **SPLIT TABLE**
Chunk 1: "| Controller | Processor |"
Chunk 2: "| Must comply | Must assist |"
→ table split mid-row, unusable for retrieval

INTACT TABLE
"| Controller | Processor |
| Must comply | Must assist |"
→ complete table in one chunk

SC — Size Compliance

Proportion of chunks within target token bounds [100–1,100]. Oversized chunks dilute embeddings; tiny chunks waste retrieval slots.

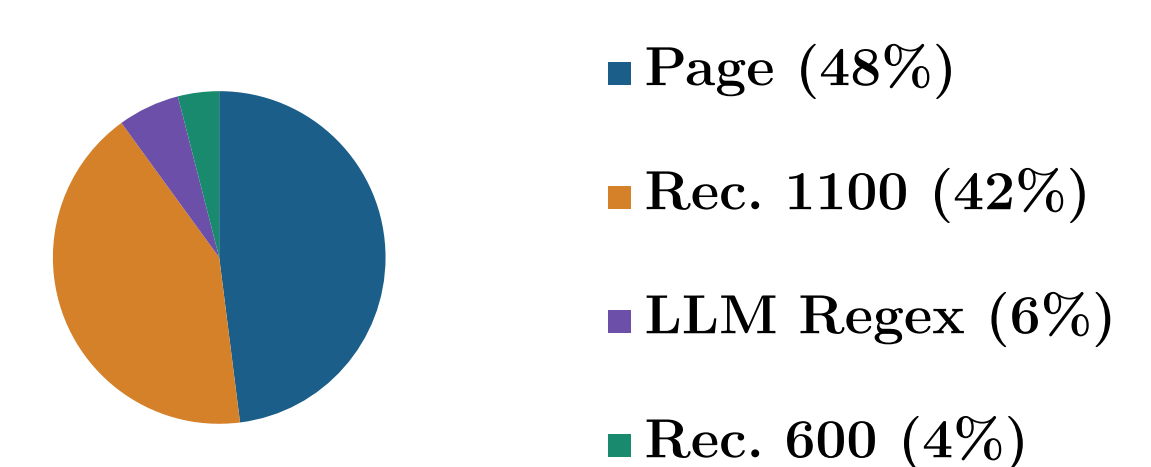


Pairwise inter-metric Spearman ρ ranges from -0.44 to 0.31 → metrics capture complementary phenomena, not a single latent factor.

Intrinsic Metrics Results

Method	RC	ICC	DCC	BI	SC	Mean
LLM Regex	98.0	70.9	82.4	98.1	99.6	89.8
Recursive 1100	99.0	66.6	89.7	98.1	100	90.7
Recursive 600	97.2	69.6	84.7	94.8	100	89.2
Page	97.2	69.2	86.4	99.9	99.9	90.5
LC recursive	96.1	65.6	88.8	95.0	97.7	88.6
Semantic	97.5	69.3	76.3	91.3	48.1	76.5
Adaptive	99.0	68.2	88.8	99.4	99.9	91.1

How Often is Each Method Selected?



RAG Performance

+33% Questions Answered
72% Mean RAG Perf.

Metric	Ours	LC Rec.	Page
Retr. Compl.	67.7	58.1	59.1
Ans. Correct.	78.0	70.1	73.3
Answered	65/99	49/99	49/99

Intrinsic gains of 0.4–2.4 pp compound ×4–5 through the pipeline. All from chunking alone.

Key Takeaways

- **Document-aware selection beats any global default.** Per-document selection delivers higher retrieval completeness, answer correctness, and answer rates: better chunks let the LLM answer more questions.
- **Always regularize chunk sizes.** Merging tiny and re-splitting oversized chunks provides consistent gains at negligible cost.
- **Balance cohesion and context.** Multi-metric scoring avoids overfitting to one dimension.

Code: github.com/ekimetrics/adaptive-chunking
Contact: jean.lelong@ekimetrics.com,
annabelle.blangero@ekimetrics.com