ARTEFACT
AI IS ABOUT PEOPLE

48th European Conference on Information Retrieval
ECIR 2026
29 March – 2 April
Delft, The Netherlands

A Systematic Analysis of Chunking Strategies for Reliable Question Answering

Sofia Bennani & Charles Moslonka
Artefact Research Center

ARTEFACT RESEARCH CENTER

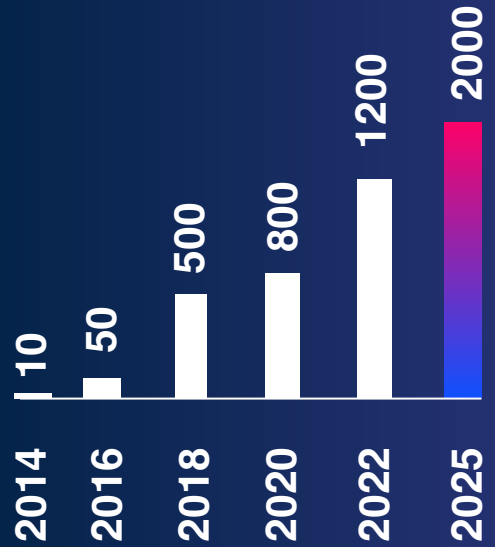
ArGiMi

FRANCE 2030

– Public –

Artefact is a world leading **data & AI consulting** partner.

#Employees



27 countries

2000 employees

36 offices



– Public –

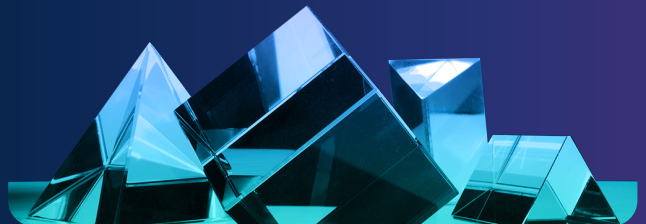
ARTEFACT

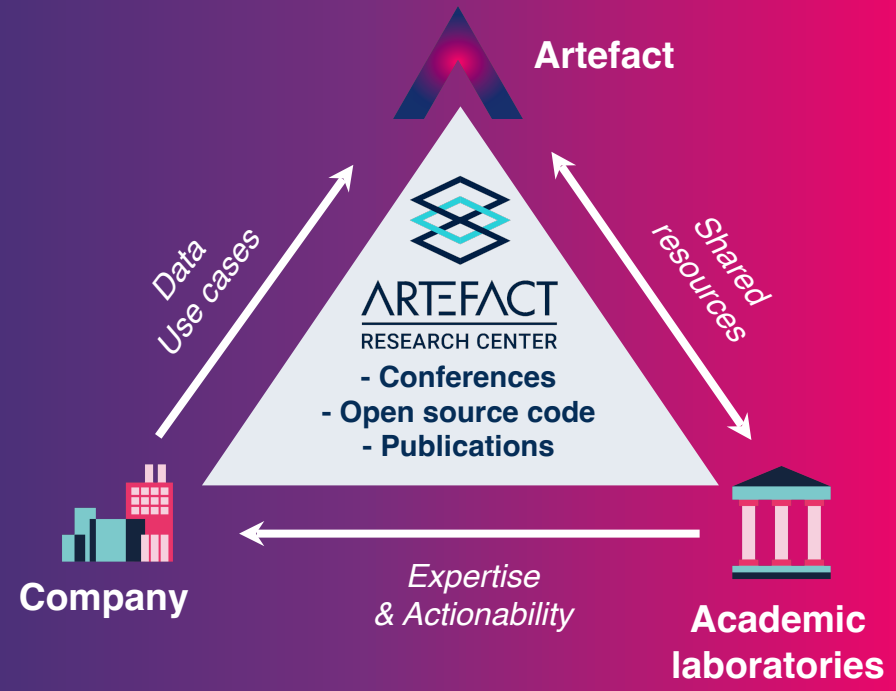
Artefact Research Center: Bridging the gap between academia and industry applications.

ARTEFACT RESEARCH CENTER

The Artefact Research Center's mission:

- A full ecosystem bridging the gap between **fundamental research** and industry tangible **applications**.
- Collaboration between Artefact and **academic researchers**, combined with PhD students from top universities and engineering & business schools.
- Research on **more robust, transparent and ethical models** to nurture **AI** business adoption.





– Public –

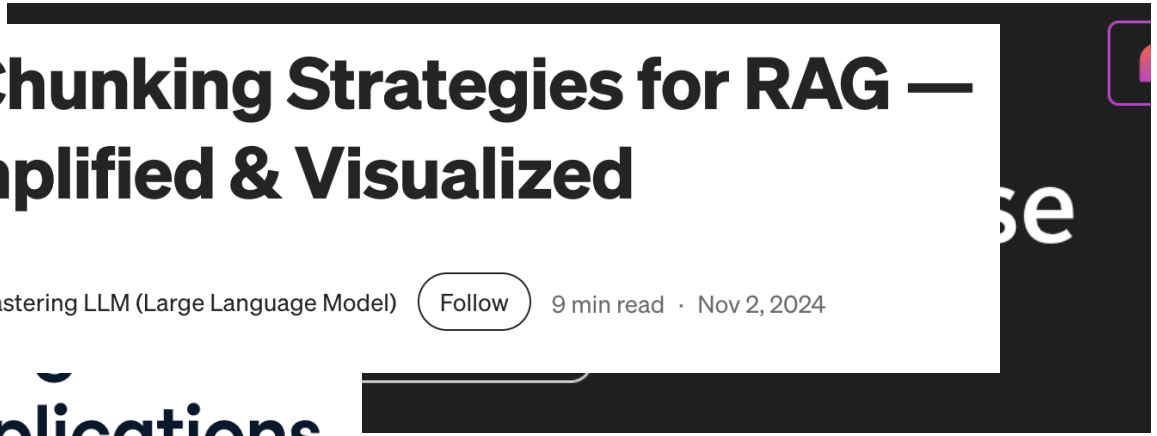
Chunking in RAG pipelines.

Saying Retrieval-Augmented Generative pipelines are (were?) popular is an understatement.

Yet chunking strategies are usually left to rule of thumb / gut feeling in Medium blog posts.

We wanted to have **rigorous starting point** for projects/client cases, and actionable findings.

- Go-to parameters for prototyping.
- Typical use-case: **QA for textual data.**
- **End-to-end** methodology.

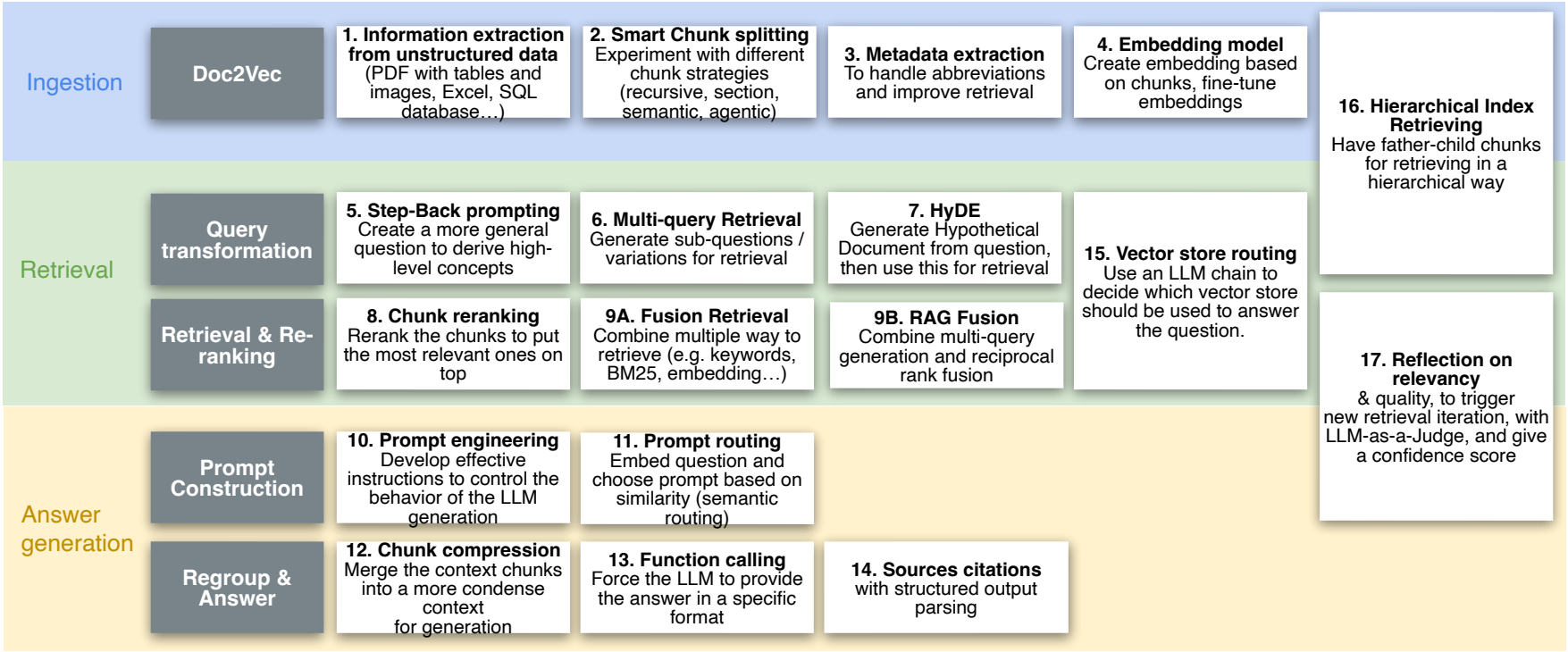




– Public –

ARTEFACT

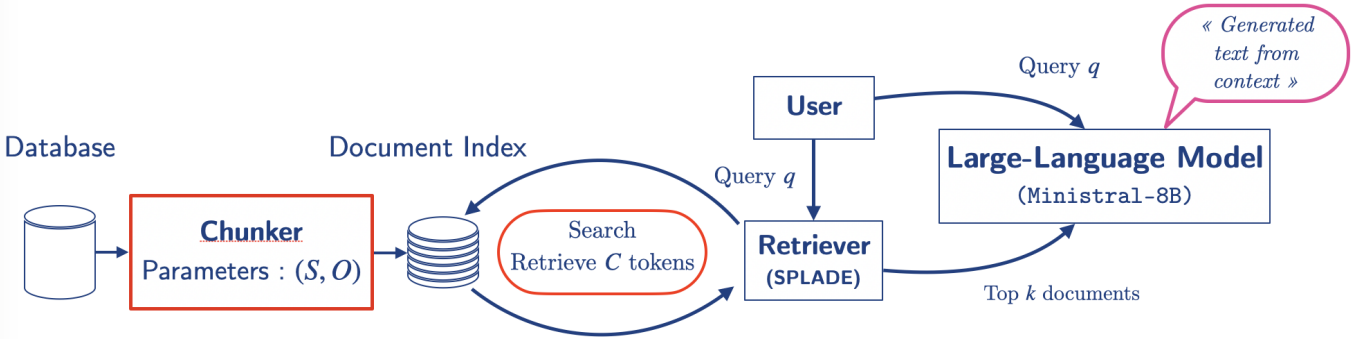
A typical production pipeline



– Public –

ARTEFACT

What we tested: Experimental set-up



Focus on chunking

We tried four different chunking methods: **Token, Sentence, Semantic, Code.**
Along with 3 parameters:

- Chunk size S
- Overlap between chunks O
- Number of retrieved tokens: Context C

Around 200 indexes

- Single-stage Sparse Learned Retriever: **SPLADE++**
- LLM for the output: `Ministral-8B`, low temperature, allowed abstention.
- Benchmarked on NQ : **2M-8M chunks** indexed
- 3 metrics : BERTScore, ExactMatch, and Abstention ratio.

– Public –

ARTEFACT

Our findings: Practical Defauts for Question Answering

Choice	Default	Rationale
Overlap O	0%	No measurable benefit; reduces cost/complexity
Chunker	Sentence	Matches semantic up to ~5k tokens; cheaper
Chunk size S	150–300	Balances recall vs. abstention
Context C (QA)	~2.5k	Avoids context cliff; boosts EM
Context C (Summ.)	~500	Maximizes semantic faithfulness
When $C > 5$k	Consider Semantic	Slight edge at very large contexts

– Public –

Overlap is mostly a waste of money

No measurable gains, for added index size.

We checked values of Overlap from 0 up to 60%.

- No noticeable differences for 10-20% for all values of chunk size S and context C.

  ΔBertScore < 5.10^(-4), ΔEM < 0.001

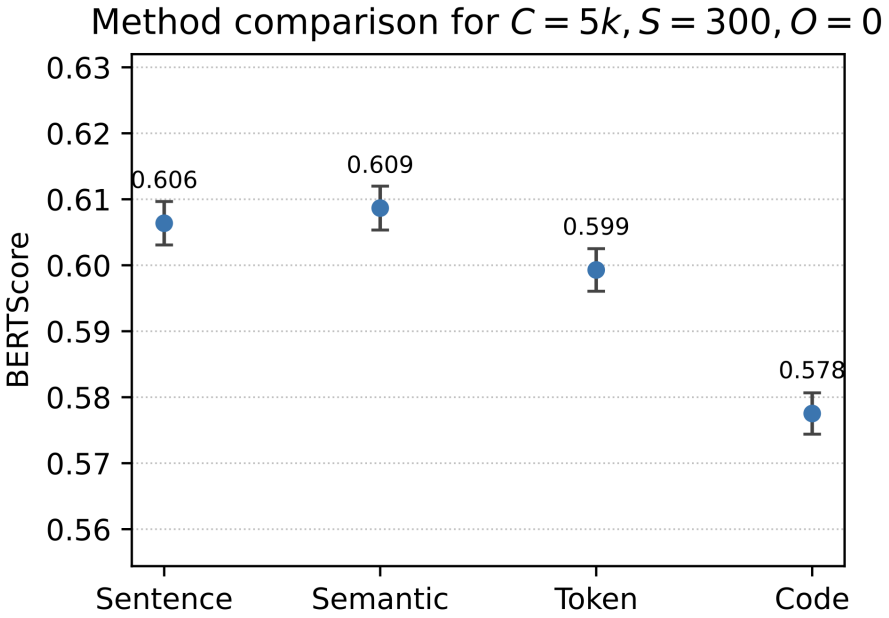
- Inflation factor I: 25% more documents for 20% Overlap. $$I = \frac{1}{1 - \text{Overlap}}$$

– Public –

Prefer Sentence Chunking

Sweet spot between performance and indexing overhead.

- Sentence ≈ Semantic performance wise.
- Sentence does not requiere an embedding model at indexing time: **saves time and money.**
- Token is noticeably worse.
- Code (markdown) is irrelevant (for text documents).

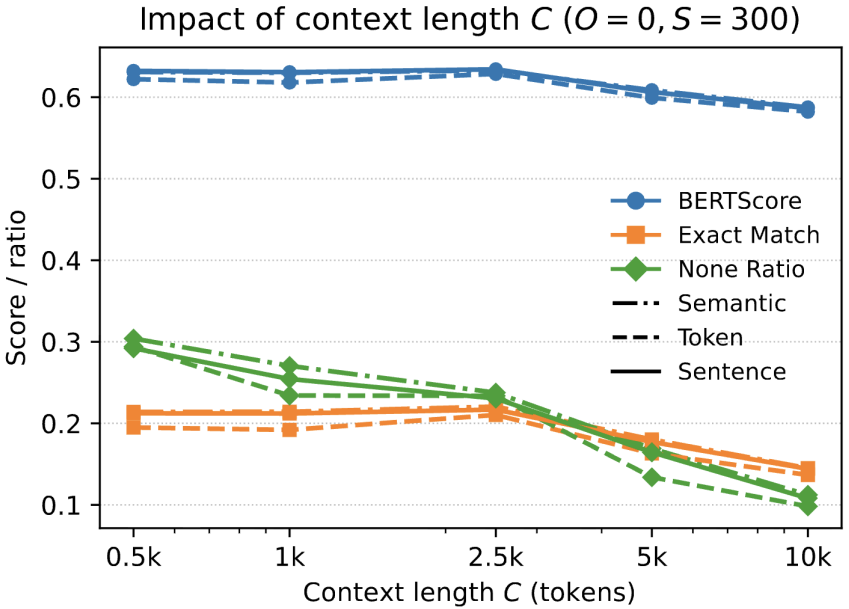


– Public –

Beware of the “context cliff”

Feeding too much information leads to worse performance.

- Large Context C = Lots of retrieved chunks.
- Abstention ratio and Exact Match plummet after 2.5k tokens.
- More context = more noise for the LLM.



– Public –

Parameter tuning

Tune parameters by task.

- **Chatbot / summarizer** (user expects fluent, faithful answers):
  Keep context small (~500 tokens)
- **Factoid QA / search assistant** (user expects precise, correct answers):
  Use moderate context (~2.5k tokens).
- **Low-abstention system** (can't afford "I don't know"):
  Increase C and decrease S.
  More small chunks = more chances to surface the right passage, but **increased hallucinations.**

  No universal "best config" — Prompt configuration matters.

– Public –

What we did not test.

Multiple limitations to acknowledge

- **No-reranker**. Design choice to isolate chunker.
- **Single-model** (open-weights) context effect may vary + fast obsolescence
- **NQ** means English Wikipedia : narrow benchmark.
- No perceived quality measure.

Yet : Practical starting point for our Data Scientists to iterate on.

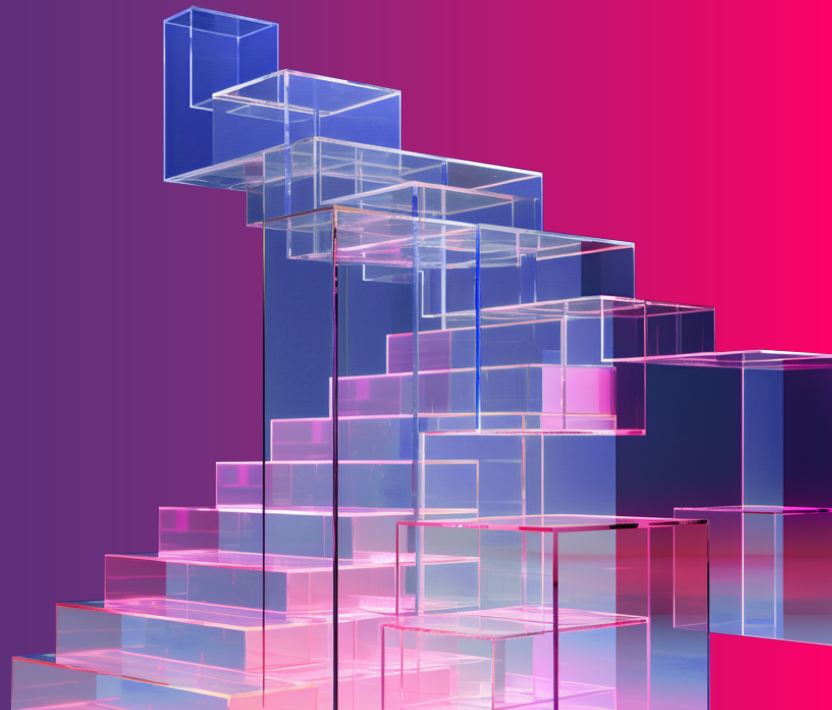
– Public –

48th European Conference on Information Retrieval

ECIR 2026

29 March – 2 April
Delft, The Netherlands

Thank you.



ARTEFACT

– Public –