

Beyond Chunk-Then-Embed

A taxonomy and evaluation of document chunking strategies for information retrieval

Yongjie Zhou, Shuai Wang, Bevan Koopman, Guido Zuccon

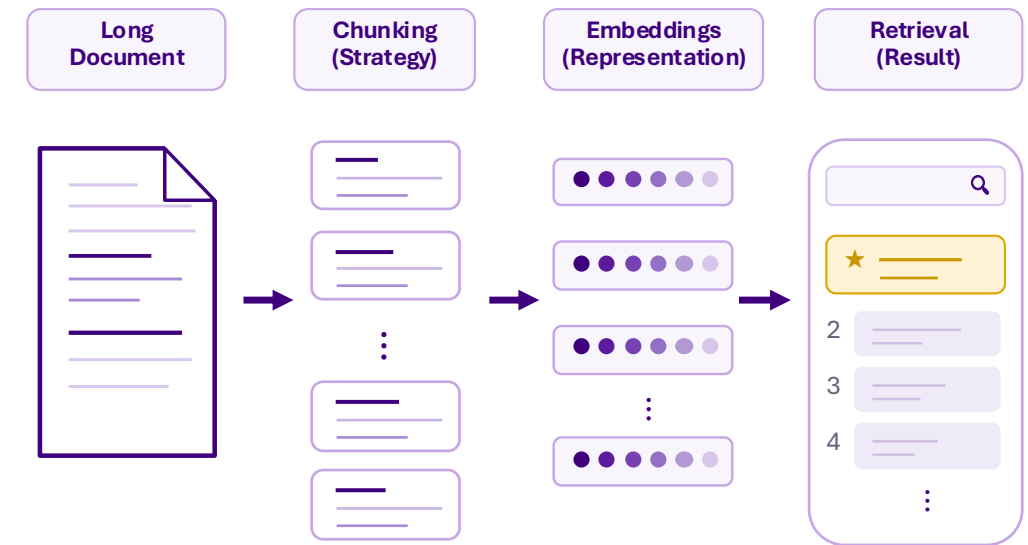
The University of Queensland

Presenter: Yongjie Zhou

yongjie.zhou@uq.edu.au



SIGIR 2026 • Melbourne, Australia



Why Chunking Decisions Matter?

Dataset: TREC-COVID

Query



what are the best
masks for preventing
infection by Covid-19?

Relevant Document (ID: 0durj95f)



"... filtration efficiencies for various fabrics when a **single layer** was used ranged from **5 to 80%** and **5 to 95%** for particle sizes of <300 nm and >300 nm, respectively, the efficiencies improved when **multiple layers** were used and when using a specific combination of different fabrics. Filtration efficiencies of the **hybrids** (such as cotton-silk, cotton-chiffon, cotton-flannel) was **>80%** (for particles <300 nm) and **>90%** (for particles >300 nm). We speculate that the enhanced performance of the hybrids is likely due to the combined effect of mechanical and electrostatic-based filtration. Cotton ... **performs better at higher weave densities** ..."

Why Chunking Decisions Matter?

Query



what are the best
masks for preventing
infection by Covid-19?

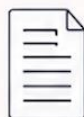
Relevant Document (ID: 0durj95f)



"... filtration efficiencies for various fabrics when a **single layer** was used ranged from **5 to 80%** and **5 to 95%** for particle sizes of <300 nm and >300 nm, respectively, the efficiencies improved when **multiple layers** were used and when using a specific combination of different fabrics. Filtration efficiencies of the **hybrids** (such as cotton-silk, cotton-chiffon, cotton-flannel) was **>80%** (for particles <300 nm) and **>90%** (for particles >300 nm). We speculate that the enhanced performance of the hybrids is likely due to the combined effect of mechanical and electrostatic-based filtration. Cotton ... **performs better at higher weave densities** ..."

Too Small

Example Chunks



"... **single layer** was used ranged from **5 to 80%** ..."

"... **5 to 95%** for particle sizes of <300 nm and >300 nm ..."

"... **multiple layers** were used ..."

"... **hybrids** (such as cotton-silk, cotton-chiffon, cotton-flannel) ..."

Retrieval Result



Fragmented evidence

Lower score



Key information is split across multiple tiny pieces.

Why Chunking Decisions Matter?

Query



what are the best
masks for preventing
infection by Covid-19?

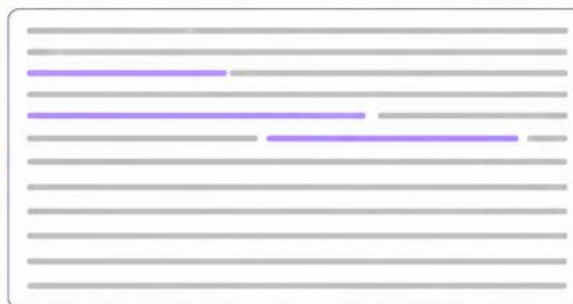
Relevant Document (ID: 0durj95f)



"... filtration efficiencies for various fabrics when a **single layer** was used ranged from **5 to 80%** and **5 to 95%** for particle sizes of <300 nm and >300 nm, respectively, the efficiencies improved when **multiple layers** were used and when using a specific combination of different fabrics. Filtration efficiencies of the **hybrids** (such as cotton-silk, cotton-chiffon, cotton-flannel) was **>80%** (for particles <300 nm) and **>90%** (for particles >300 nm). We speculate that the enhanced performance of the hybrids is likely due to the combined effect of mechanical and electrostatic-based filtration. Cotton ... **performs better at higher weave densities** ..."

Too Large

Example Chunk



Retrieval Result



Relevant evidence
buried in noise

Lower score



Irrelevant content overwhelms the evidence.

Why Chunking Decisions Matter?

Query



what are the best
masks for preventing
infection by Covid-19?

Relevant Document (ID: 0durj95f)



"... filtration efficiencies for various fabrics when a **single layer** was used ranged from **5 to 80%** and **5 to 95%** for particle sizes of <300 nm and >300 nm, respectively, the efficiencies improved when **multiple layers** were used and when using a specific combination of different fabrics. Filtration efficiencies of the **hybrids** (such as cotton-silk, cotton-chiffon, cotton-flannel) was **>80%** (for particles <300 nm) and **>90%** (for particles >300 nm). We speculate that the enhanced performance of the hybrids is likely due to the combined effect of mechanical and electrostatic-based filtration. Cotton ... **performs better at higher weave densities** ..."

Good Chunk

Example Chunk



"... a **single layer** was used ranged from **5 to 80%** and **5 to 95%** for particle sizes of <300 nm and >300 nm ... **multiple layers** ... **hybrids** ... **>80%** (for particles <300 nm) and **>90%** (for particles >300 nm) ... Cotton ... **performs better at higher weave densities** ..."



Retrieval Result



Focused, complete
evidence

Higher score

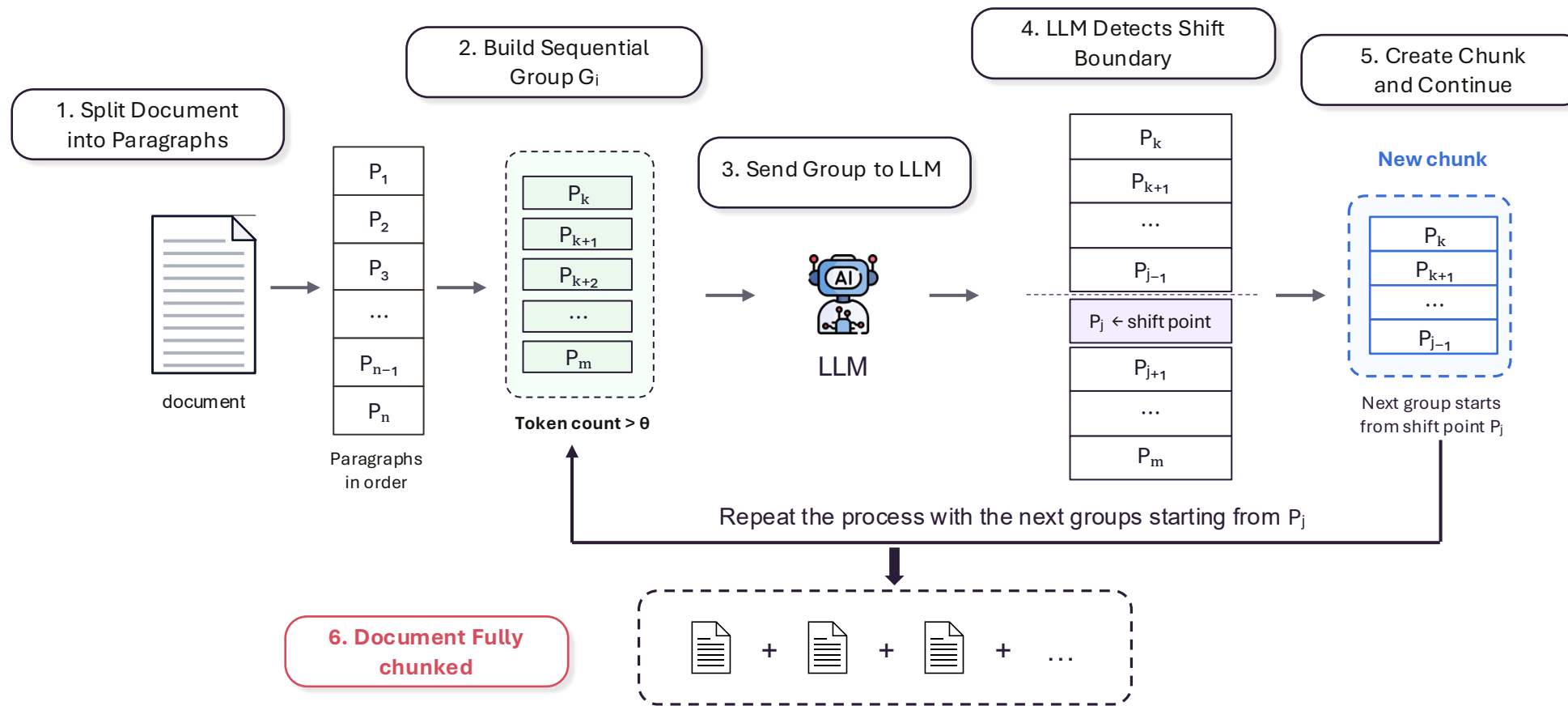


Good chunk captures the complete evidence.

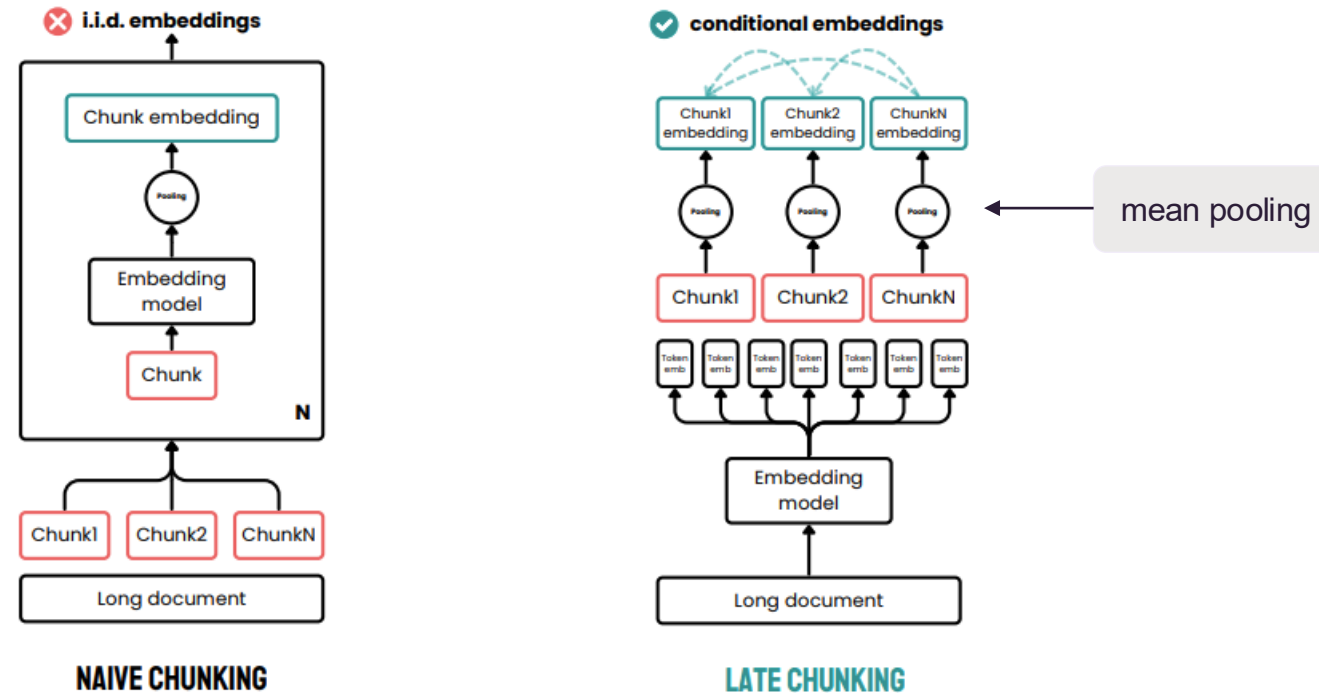


The right chunking strategy preserves context, keeps evidence intact, and improves retrieval quality.

Lumberchunker: Long-form narrative document segmentation. [1]



Late chunking: contextual chunk embeddings using long-context embedding models. [2]



- LumberChunker focuses on how to **segment** .
- Late Chunking focuses on how to **encode**.
- All are evaluated **in isolation**, using **different tasks**, **datasets**, and **model**.



We need a unified evaluation framework. That is what we aim to build.

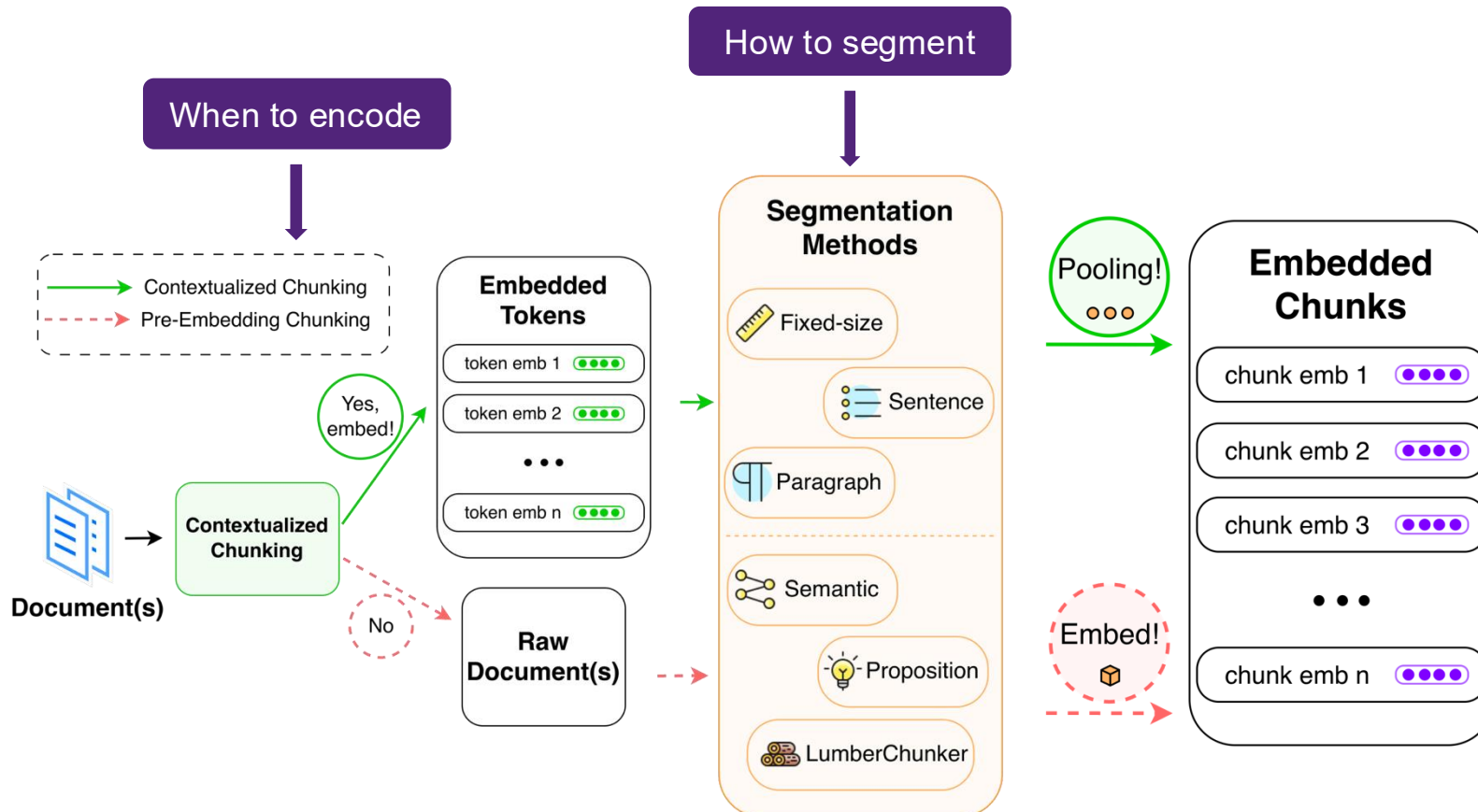
Our Work: Taxonomy and Evaluation

1 Embedding order

decides whether chunks are encoded independently or in document context

2 Segmentation method

decides where boundaries are.



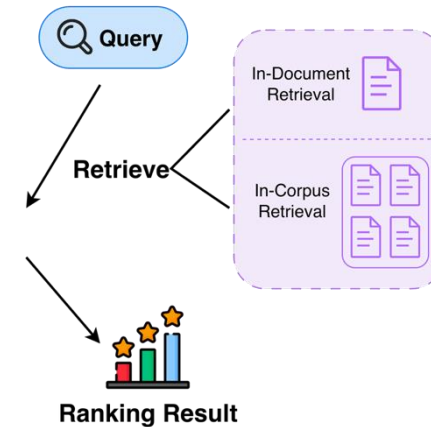
Evaluation Tasks



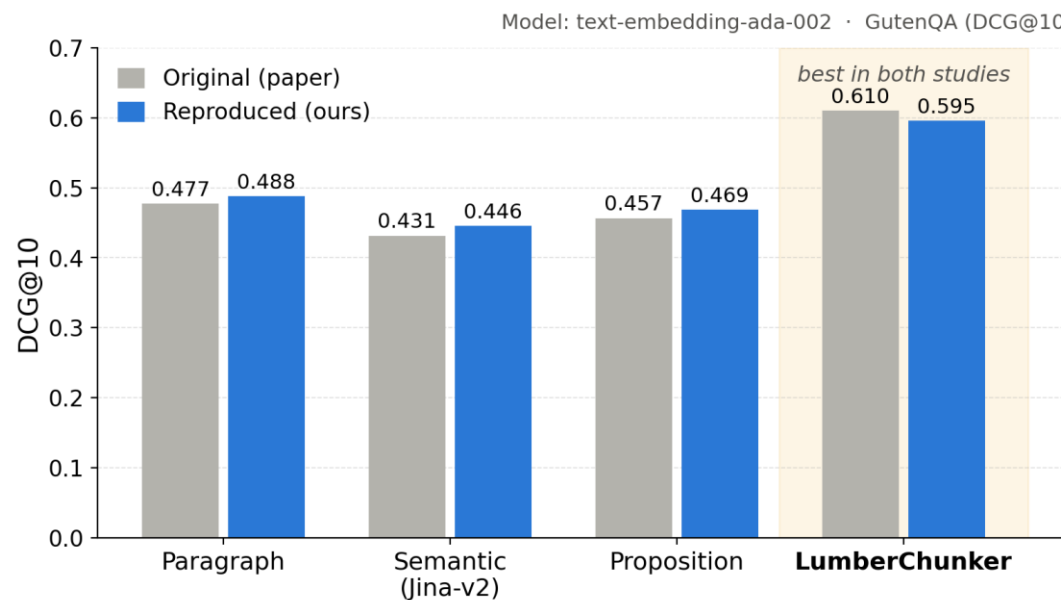
In-document retrieval



In-corpora retrieval

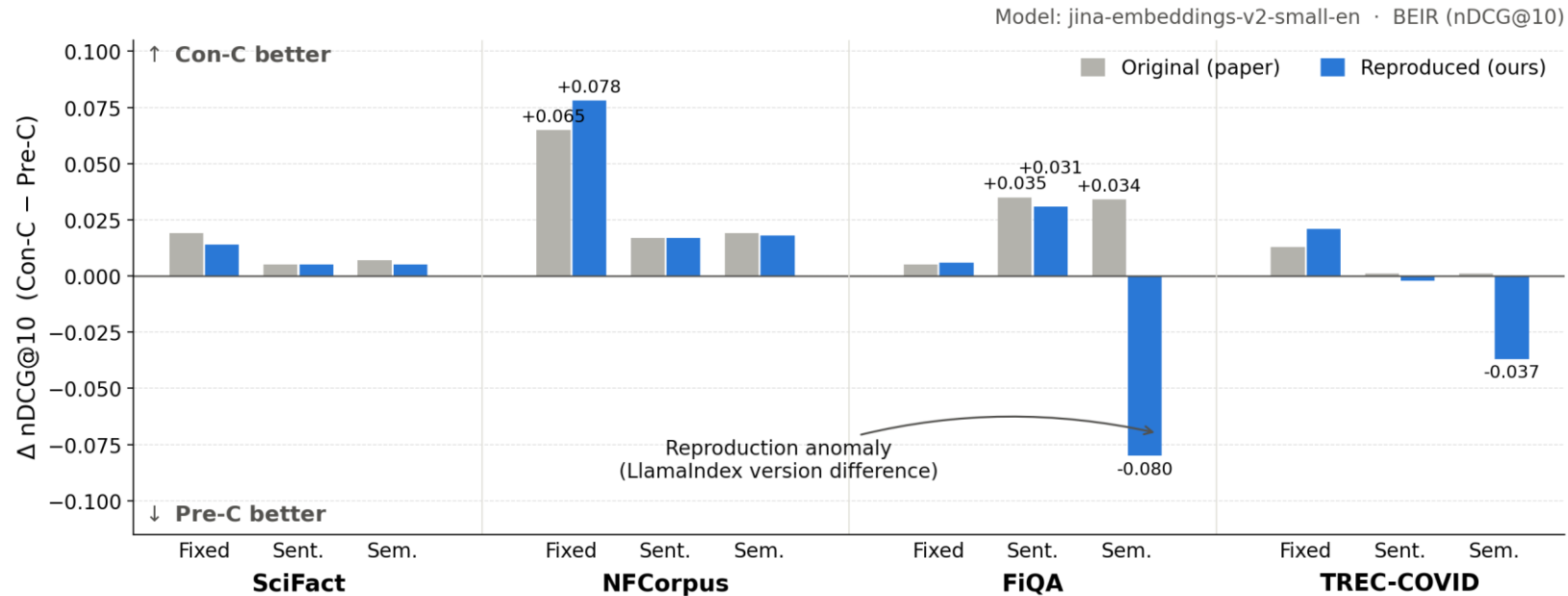


Original Claim 1: LumberChuker wins in-document retrieval



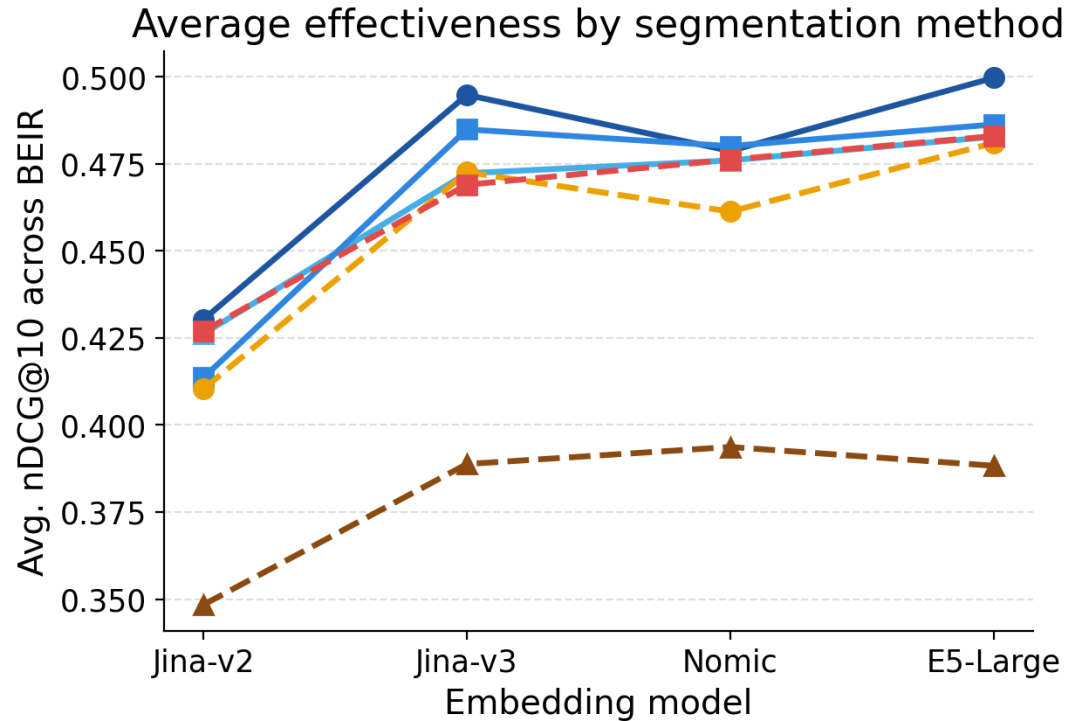
Reproduced Overall close alignment with the original paper

Original Claim 2: Late chunking helps in-corpora retrieval



Mostly reproduced Late Chunking shows consistent gains

RQ1: Impact of Segmentation Methods on In-Corpus Retrieval



- Best average methods are **structure-based** across all four models.
- **Proposition** segmentation consistently is the worst.



**Practitioner rule: Start with simple structure
(Paragraph or fixed-size) for in-corpus retrieval.**



Higher accuracy



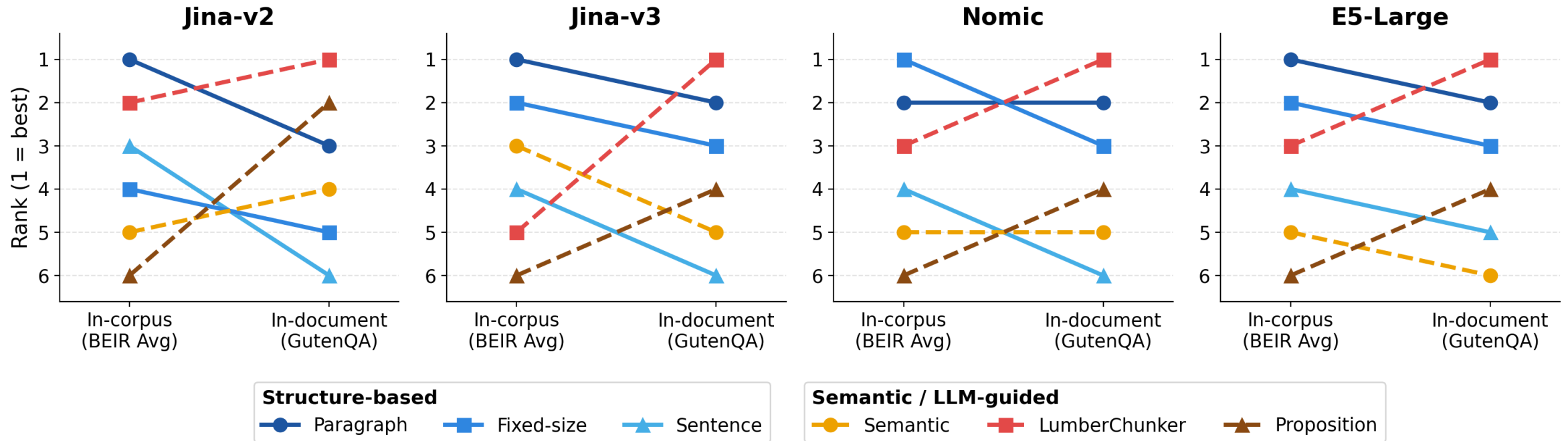
Lower cost



Low latency

RQ2: Do Segmentation Methods Generalize Across Retrieval Tasks?

Segmentation-method ranking reverses between retrieval tasks

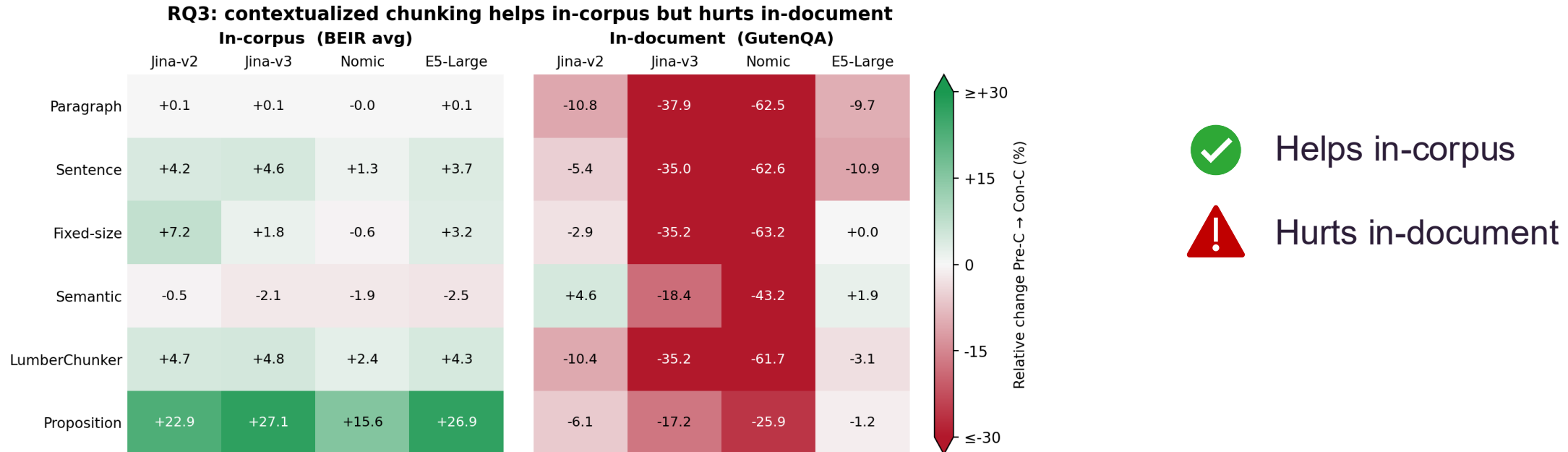


The best segmentation method changes when retrieval within a single long document.



Task matters: structure-based methods win in-corpora; LumberChunker wins in-document.

RQ3: Effectiveness of Contextualized Chunking



- In-corpus retrieval: **mostly positive** +2.78% to 6.41%, Proposition gains most +15.62% to +27.11%
- In-document retrieval: **mostly negative** -3.83% to -53.17%



Takeaway: Contextualized chunking is task-dependent. Use it for in-corpus retrieval; avoid it for in-document retrieval.



Three simple rules:

- 1 Searching across documents? Use Paragraph or Fixed-size
- 2 Searching within one long document? Use LumberChunker
- 3 Using contextualized chunking? Use for in-corpus; avoid for in-document



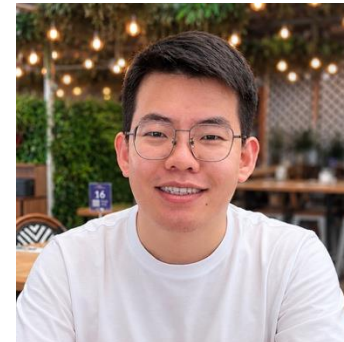
Code

<https://github.com/ielab/Reproduce-chunking-2026>



Paper

<https://dl.acm.org/doi/10.1145/3805712.3808575>



Presenter: Yongjie Zhou

yongjie.zhou@uq.edu.au

<https://zhouyongjie.github.io/>