

Beyond Abstracts

Learning Scientific Paper Embeddings from Full-Text Windows

Younes Djemmal, Oloruntobi Olutola, Kim Gerdes · Questel · Université Paris-Saclay, LISN · Model presentation · IR 2, Session 1 · Nov 6, 2026

The problem: papers are longer than their abstracts

- Citation-trained paper embeddings (SPECTER, SciNCL, SPECTER2) read **title + abstract only**: about 250 words
- A paper body has on average **3,900 words** in 9.5 sections
- Many real queries are **passages from a body**: a method paragraph, a result, a claim to check against prior art
- Question: can an encoder **learn from the full text** and still be good at abstracts?

What the spring pipeline does, and what this paper changes

Spring (class 3 challenge)

- an **off-the-shelf** bi-encoder (MiniLM, SPECTER2)
- encodes **title + abstract**
- queries are short or abstract-like
- citations used by SPECTER, but only abstract-to-abstract

This paper

- **trains** the bi-encoder (SciBERT, ModernBERT)
- on **windows of the body text**
- queries can be **body passages**
- citation pairs + 20 % **same-paper** pairs (an abstract and a window of the same paper)

Data and method

- **S2Full**: peS2o full texts + the Semantic Scholar citation graph: **2.52M papers, 4.76M citation pairs, 19 domains**; test: 10,022 queries over 63,095 unseen papers
- Two kinds of **views** of a paper: title + abstract (TA), or a **body window** (a section header and the text that follows)
- **Positive pairs**: a paper and a paper it cites, each in a random view, + 20 % same-paper pairs
- **Loss**: InfoNCE with in-batch negatives (batch 2,048): the loss from the recap
- **Control** `ta_ta`: same training, abstracts only

Filtering stage	Papers	Pairs
Semantic Scholar metadata	6,282,865	—
+ venue filter	5,193,622	—
+ fields-of-study filter	4,545,158	—
Raw citation edges (both endpoints valid)	—	27,310,848
+ bidirectional dedup + self-citation removal	—	22,171,946
+ high-degree filter ($>$ 200 pairs)	—	20,628,139
+ text quality filter	3,602,952	18,011,917
+ domain cap $\rightarrow$ training set	2,516,974	4,762,877
Test split	63,095	55,030

Table 1: the S2Full funnel (Djemmal, Olutola & Gerdes 2026)

Evaluation: three scenarios

Scenario	Query	Corpus	Relevant
TA → TA	title + abstract	titles + abstracts	the papers it cites or is cited by
Body → TA	one body window	titles + abstracts	same
Body → Body	one body window	one body window per paper	same

- Metric: **nDCG@10**; a paper never retrieves itself
- Baseline: **SPECTER2** (proximity adapter)

Main result

Model	TA→TA	Body→TA	Body→Body
SB both_random	0.518	0.439	0.391
SB anchor_r._pos_ta	0.528	0.441	0.366
SB no_ta_ta	0.517	0.439	0.389
SB ta_ta	0.537	0.406	0.304
MB both_random	0.535	**0.485**	**0.457**
MB ta_ta	**0.544**	0.454	0.358
SPECTER2	0.419	0.273	0.192

Table 4: nDCG@10 on the test set (windowed)

- **Body → Body**: 0.457 vs **0.192** for SPECTER2
- **Body → TA**: 0.485 vs **0.273**
- **TA → TA**: 0.544 (abstract-only control) vs **0.419**
- Body training does not cost much on abstracts: 0.535 vs 0.544

Which part did the work?

Controlled comparison, SciBERT, same framework and data:

	TA→TA	Body→TA	Body→Body
trained on body windows (`both_random`)	0.518	**0.439**	**0.391**
trained on abstracts only (`ta_ta`)	**0.537**	0.406	0.304
difference	−1.9	+3.3	+8.7

- The body text helps **body queries**, most for Body → Body
- But `ta_ta` already beats SPECTER2 by 13 points on Body → TA: **most of the gap comes from the training framework and the data**, not from the body text

Our testbed row

nDCG@10	BM25	MiniLM	Beyond Abstracts	Δ vs BM25 [95 % CI]	p
SciFact · 300 claims	0.676	0.646	**0.717**	**+0.040** [+0.007, +0.074]	0.02
NFCorpus · 323 queries	**0.321**	0.316	0.313	−0.008 [−0.031, +0.015]	0.50
ArguAna · 1,406 arguments	**0.516**	0.503	0.461	**−0.055** [−0.074, −0.036]	< 0.001

- Beats BM25 where it was trained, **scientific claims → abstracts**, but only level with BGE-small (0.710) and RRF (0.718)
- No gain on lay medical queries; a clear **loss on counter-arguments**, which citations never taught it
- Scored like yours: pytrec_eval, paired t-test over queries, bootstrap CI

What did not work, and what we did not test

- **On the testbed**: no gain on NFCorpus, **−0.055 on ArguAna**: it learned “cites the same topic”, not “answers this query”
- **No significance tests** in the paper, one run per configuration
- **One baseline**: SPECTER2, not retrained on our data; no BM25, no hybrid, no recent general embedder
- ModernBERT trained for fewer epochs than SciBERT
- Citations are a **proxy** for relevance; no real user queries
- Not ablated: the 20 % self-alignment, the projection head, the context length

Demo: three queries on SciFact

Hero: *"The amount of publicly available DNA data doubles every 10 years."* Labelled relevant: *The EMBL nucleotide sequence database*: BM25 rank 631, Beyond Abstracts **rank 1**

Failure: *"Venules have a thinner or absent smooth layer compared to arterioles."* Labelled relevant: an abstract on fibroblast cultures that lists venules once: BM25 **rank 1**, Beyond Abstracts rank 81

Wildcard: a claim from the room (backup: *"Aspirin reduces the risk of colorectal cancer."*, then the same claim without its keywords: *"Taking a common painkiller every day makes bowel tumours less likely."*)

What discussant A, the auditor, should ask

1. **Validation vs test.** Fig. 5 gives TA → TA ≈ 0.20 on validation, Table 4 gives 0.52–0.54 on test. Appendix C explains why Body → TA validation is inflated, not this gap. Are the two sets built the same way, and was anything chosen on validation?
2. **Noise.** One run per configuration, no significance test, and the largest per-domain gains (Art +31.8, History +14.1) come from domains with **10 test queries**. Which differences survive a paired test?

And the kind of detail an auditor notices: Table 6 bolds 0.163 for `no_ta_ta`, but `both_random` has 0.162 with ModernBERT.

What discussant B, the skeptic, should ask

1. **Is it the body text?** Your own control `ta_ta` already beats SPECTER2 by 13 points on Body → TA. Is the result about the body text, or about training on 4.76M citation pairs with a batch of 2,048? And why no BM25 and no recent general embedder?
2. **Is it relevance?** Citations stand in for relevance. Would an examiner holding a methods paragraph accept the cited papers as the answers? And Body → Body means indexing windows instead of abstracts: at what cost?