SciEmbed: Citation-Context Supervision for Scientific Document Embeddings

Jonas Wilinski
Hamburg University of Technology
jonas.wilinski@tuhh.de

Michael Färber
TU Dresden, ScaDS.AI, Dresden
michael.faerber@tu-dresden.de

Abstract

Open scientific document embedders are typically trained on binary citation edges between paper title-abstract pairs. This signal discards the citation context sentence: a domain-expert description of the cited paper, written by the citing author. We use these sentences as the primary contrastive supervision signal to train **SciEmbed**, an open 149M ModernBERT-base encoder. On the official SciRepEval benchmark it reaches 66.85 over three seeds, surpassing prior open scientific embedders such as SPECTER2 and SciNCL. To isolate what the supervision source contributes, we compare against IBM Granite R2, a same-size ModernBERT peer trained on general retrieval data. Using only open citation-context data, SciEmbed matches Granite R2 to within 0.6 on the SciRepEval macro overall while specialising in the opposite direction: it leads regression tasks tied to paper impact by 2.3 points, where Granite R2 instead leads classification, proximity, and search. A 28-configuration sweep over recipes and alternative supervision signals does not improve on this result, consistent with the specialisation tracking the supervision source rather than the training recipe. Training at longer contexts yields monotonic gains on a body-fact retrieval probe within our recipe. We release the model, training pipeline, and evaluation harness as a fully open, citation-context-supervised baseline with a distinct, reproducible specialisation pattern.

1 Introduction

Recent years have witnessed a rapid expansion of embedding models for scientific literature. However, existing approaches remain bound to binary citation graphs, LLM-generated data or rigid task-specific fine-tuning (Singh et al., 2023; Brinner and Zarrieß, 2025; Ostendorff et al., 2022; Mysore et al., 2022). In the case of citation-based specialized models, we often see the use of binary edges between citing and cited title+abstract pairs. This edge representation discards the sentence in which the citing paper explains *why* the prior work is referenced. A sentence such as "We extend the framework of Smith et al. to handle multi-modal scientific text" functions as a natural-language query for the cited paper; it is written by a domain expert who has read both works. We therefore treat the citation context sentence as the primary supervision signal. Most of these models additionally use a rather outdated 2019 SciBERT (Beltagy et al., 2019) backbone. The 512-token ceiling discards $> 90\%$ of an average full-text paper (Lo et al., 2020) and rules out the long-context retrieval scenarios now standard in modern IR benchmarks (Zhu et al., 2024).

Two recent developments now make this signal usable at scale. The Semantic Scholar Academic Graph (S2AG) (Kinney et al., 2023) contains roughly one billion citation context sentences together with citation intent labels (Cohan et al., 2019). Separately, ModernBERT (Warner et al., 2025) provides an 8,192-token encoder built on RoPE (Su et al., 2024) and Flash Attention 2 (Dao, 2024), pretrained on 2T tokens and validated for biomedical continued pretraining (Sounack et al., 2025). To our knowledge the two have not yet been combined for open scientific document embeddings.

We train SciEmbed, an open 149M ModernBERT-base encoder, in two stages: domain-adaptive MLM on scientific full text, followed by contrastive fine-tuning on the S2AG citation graph. The contrastive supervision takes the citation context sentence as the load-bearing information, not solely the citation edge (Figure 1). To isolate what this supervision source contributes at fixed architecture, we compare against IBM Granite Embedding R2 (Awasthy et al., 2025), which shares our backbone and recipe family but trains on general retrieval data.

SciEmbed-FULL reaches 66.85 ± 0.38 on the official 22-task SciRepEval benchmark (Singh et al.,

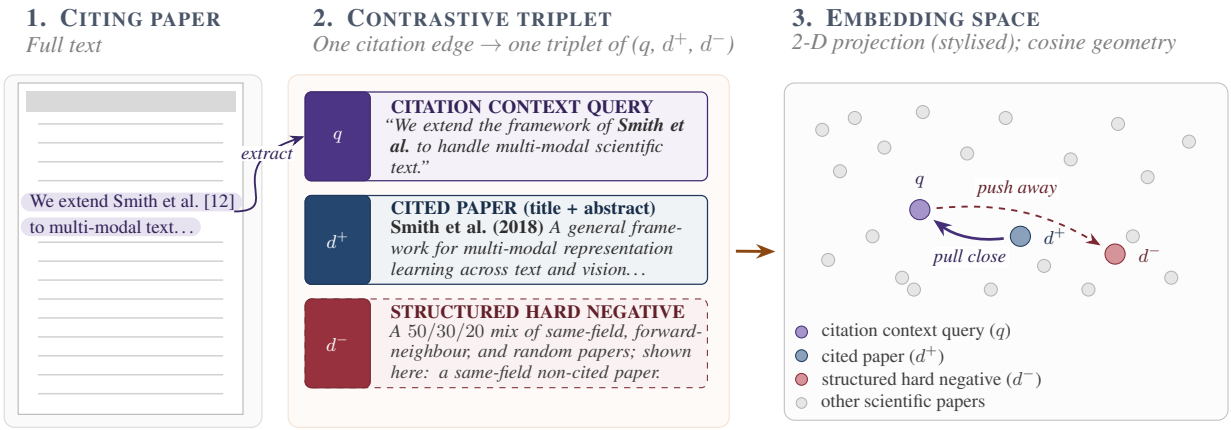


Figure 1: A citing paper contains a citation context sentence: a domain-expert, natural-language description of the cited work. Each citation edge yields one contrastive triplet: the context sentence as query q, the cited paper's title+abstract as positive d^+, and a structured hard negative d^- (Section 3.3). SciEmbed (149M-parameter ModernBERT-base, 8K context) is trained to pull d^+ close to q and push d^- away.

2023) over three seeds, surpassing prior open scientific embedders and matching the same-size Granite R2 to within 0.6. This small gap hides the more informative question of *where* the two specialise.

Contributions. We make four primary contributions to the field of scientific document representation:

- We train and openly release SciEmbed, a 149M ModernBERT-base scientific embedder supervised primarily by citation context sentences, together with its training pipeline and evaluation harness – a fully open, citation-context-supervised baseline with a distinct, reproducible specialisation pattern.[1]
- We isolate the supervision source at fixed architecture through a controlled comparison with the same-size Granite R2, finding that the two models specialise on opposite SciRepEval buckets, a mirror image no other peer reproduces (Section 5.1).
- We characterise recipe-space saturation: within the tested configuration space, a 28-configuration recipe sweep never beats citation-context supervision and Stage-1 domain-adaptive MLM contributes ≤ 0.1 on SciRepEval, consistent with the result being bounded by the supervision source rather than the recipe (Sections 5.3 and 5.2).
- We probe long-context behaviour with a Body-Fact Retrieval diagnostic and LongEmbed, and find monotonic gains in Stage-2 context length within the recipe (Section 5.4).

[1]Code, training pipeline, and evaluation harness: https://github.com/J0nasW/SciEmbed-release. Model checkpoints: https://huggingface.co/J0nasW.

2 Related Work

Prior work on scientific document embeddings clusters into two families, and the cell at their intersection is what motivates SciEmbed.

Citation-edge supervision on SciBERT-512. The widely used open scientific embedders are SPECTER (Cohan et al., 2020), SPECTER2 (Singh et al., 2023), SciNCL (Ostendorff et al., 2022), ASPIRE (Mysore et al., 2022), and SciMult (Zhang et al., 2023). They all build on SciBERT (Beltagy et al., 2019) or a comparable BERT-base backbone capped at 512 tokens, and supervise contrastively on *binary citation edges*: each citation between two papers becomes a positive (title+abstract, title+abstract) pair. Variants of the recipe add task-specific adapters (SPECTER2's four `[CLF]/[QRY]/[RGN]/[PRX]` LoRAs), graph-proximity positives (SciNCL), or multi-task instruction routing (SciMult), but the underlying supervision is the same edge signal. The sentence in which the citing paper actually describes the cited work, the *citation context*, is discarded; so is any input above the 512-token cap. Citation-intent taxonomies (Jurgens et al., 2018; Cohan et al., 2019; Lauscher et al., 2022; Wright and Augenstein, 2021) and citation-context retrieval (Ritchie et al., 2008) have been studied as their own task family, but to our knowledge they have not been used as the primary supervision signal for an open, general-purpose scientific document embedder.

Replacing edges with derived natural-language queries. Two recent open scientific embedders try to mitigate the same semantic dilution by replacing the citation edge with a richer derived query: SemCSE (Brinner and Zarrieß, 2025) (183M) uses LLM-generated abstract summaries, and CoSAEmb (Singh and Singh, 2024) uses aspect-conditioned section embeddings from discourse structure. Both raise signal density above edges, but the query is generated post-hoc rather than authored *in situ* by a domain expert who has read both papers. SciEmbed's citation context sentence is the same object class, but it already exists in S2AG at no extraction cost; we re-evaluate SemCSE in Table 1 and find the same pattern as with Granite R2 (Section 5.1).

General-retrieval supervision on long-context backbones. A parallel line of work has pushed general-purpose embedders to long context: BGE (Xiao et al., 2024; Chen et al., 2024), E5 (Wang et al., 2022), GTE (Li et al., 2023), Nomic Embed (Nussbaum et al., 2025; Nomic AI, 2025), Jina v3 (Sturua et al., 2025), and at decoder scale NV-Embed-v2 (Lee et al., 2025a) and Qwen3-Embedding-8B (Zhang et al., 2025). These models train on web-scale general-retrieval data (often MS-MARCO and synthetic query-document pairs) on 8K-context backbones; they do not use citation supervision and do not target SciRepEval. The query side of these pairs is a generic search intent, not the domain-expert sentence relating two scientific papers that citation-context supervision provides. Within this family, IBM Granite Embedding R2 (Awasthy et al., 2025) (149M, ModernBERT, 8K; August 2025) is the most precise architectural match for SciEmbed, sharing our backbone and a broadly comparable contrastive-finetuning recipe family but trained on general enterprise retrieval data. It is the matched-architecture peer used throughout the paper to isolate the supervision-source contribution at fixed architecture.

The unexploited cell, and SciEmbed's position. Open *scientific* embedders trained with citation supervision remain pinned to 512-token SciBERT, and open *long-context* ModernBERT embedders target general English / enterprise retrieval (BGE, Nomic, Granite R2) or biomedical-only similarity (Clinical/BioClinical ModernBERT (Lee et al., 2025b; Sounack et al., 2025; NeuML, 2025)) rather than the multi-domain scientific corpus. The cell "citation-context supervision on a long-context backbone" is, to our knowledge, unexploited among open models we evaluated. SciEmbed sits in this cell: a 149M ModernBERT-base encoder trained with S2AG citation context sentences as the primary contrastive supervision, evaluated end-to-end on the official 22-task SciRepEval suite plus LongEmbed and a Body-Fact Retrieval diagnostic probe. The full landscape table (parameter counts, context lengths, backbones, and scientific-supervision tags across open peers) is in Appendix A.2 (Table 2).

3 Methods

SciEmbed is trained in two stages: domain-adaptive masked language model (MLM) pretraining on scientific full text (Section 3.2), followed by multi-signal contrastive fine-tuning (Section 3.3). Figure 1 shows the Stage 2 contrastive setup end-to-end.

3.1 Base Model

Our backbone is ModernBERT-base (Warner et al., 2025): a 149M-parameter encoder with an 8,192-token context window, pretrained on 2T tokens of web, code, and scientific text. We prefer ModernBERT to SciBERT (Beltagy et al., 2019) because its context window is 16× larger (8,192 vs. 512 tokens, enabling full-text input) and its pretraining corpus is substantially larger (2T vs. 3.2B tokens).

3.2 Stage 1: Domain-Adaptive Pretraining (DAPT)

We initially trained SciEmbed in two stages, but the controlled ablation in Section 5.2 shows that, once the Stage 2 contrastive signal is in place, the marginal contribution of Stage 1 to SciRepEval is within noise of the contrastive-only variant, and the same picture holds at the 8K Stage-2 cap on LongEmbed and BFR (Section 5.4). We summarise the recipe below for reproducibility and treat Stage 2 as the primary contributor throughout; practitioners reproducing SciEmbed at a comparable budget can start Stage 2 directly from raw ModernBERT-base (the NoDAPT-CTX variant), though this contribution is budget-conditional (Limitations).

We continue masked language modelling on 13.4M filtered scientific full-text papers (13.9M before length filtering; PMC-OAS, S2ORC (Lo et al., 2020), arXiv, peS2o (Soldaini et al., 2024)), with ModernBERT's BPE tokenizer at 8,192-token

SciEmbed (Signal B, citation-context supervision)

q (citation context, written by the citing author): "Inspired by [9], we conducted a more extensive and systematic study of assembling several CNN-related techniques into a single network."

d^+ *(cited paper title + abstract): Bag of Tricks for Image Classification with Convolutional Neural Networks.* "Much of the recent progress made in image classification research can be credited to training procedure refinements, such as changes in data augmentations and optimization methods. . . "

Granite R2-style (general-retrieval supervision, Natural Questions)

q (Google search query): "when was the first test tube baby born"

d^+ *(matching Wikipedia passage): Louise Brown.* "Louise Joy Brown (born 25 July 1978) is an English woman known for being the first human to have been born after conception by in vitro fertilisation, or IVF."

Figure 2: SciEmbed's query is a domain-expert sentence from the citing author that explicitly motivates the cited contribution; the Granite-style query is a short user information need with a Wikipedia answer passage. The two distributions are visibly different. Citation-context example is a real Signal-B pair from our training pool; the Granite-style row is a real Natural Questions training pair, one of the three public retrieval datasets that the Granite R2 technical report (Awasthy et al., 2025) names as in-domain training data (NQ, HotpotQA, FEVER).

chunks (128-token stride). Training optimises the standard 30%-masked log-likelihood with AdamW (lr 2×10^{-5}, weight decay 0.01, 2,000 warmup steps, cosine schedule, BF16) on 4×H200 GPUs via MosaicML Composer (The Mosaic ML Team, 2021) for 16.1 h (8.3B tokens). The cleaned corpus construction, masking schedule, full hyperparameters, and the training-loss curve are in Appendix B.1.

3.3 Stage 2: Multi-Signal Contrastive Fine-Tuning

We train the Stage-1 checkpoint with two contrastive signals from the S2AG citation graph; the construction is illustrated in Figure 1 and a real example pair next to a general-retrieval counterpart is shown in Figure 2. Ablation/probe variants are detailed in Appendix B.3. **Signal A (citation edge triplets)** follows the SPECTER recipe (Cohan et al., 2020): each example pairs `search_query:` "*citing title. citing abstract*" with `search_document:` "*cited title. cited abstract*". A two-tier sampler draws 16M high-confidence triplets from the S2AG influential-citation pool ($\sim$ 140M edges) and 34M from the general non-influential pool ($\sim$ 2.8B edges, with $\geq$ 5 citations on both sides), yielding $\sim$ 50M deduplicated triplets. **Signal B (citation context queries)** pairs an S2AG citation context sentence (`search_query:` *citation sentence*) with the cited paper (`search_document:` *cited title. cited abstract*); S2AG exposes $\sim$ 1B such sentences, of which we sample 20M from the influential tier after filtering to 50–1,000 characters, $\geq$ 3 words, and $\geq$ 0.5 alphabetic ratio. **Signal C (intent-conditioned, discarded)** extends Signal B with S2AG intent prefixes (`cite_background/method/result`); it did not improve over A+B on current benchmarks and is discussed in Appendix B.3. **Signal D (structured hard negatives)** augments the in-batch negatives with one explicit structured hard negative per anchor, sampled as a 50/30/20 mix of same-field non-cited papers, forward-neighbour papers (cited by d^+ but not by the anchor, the hard-negative definition of SPECTER (Cohan et al., 2020)), and random papers; the same-field component uses the S2AG `s2fieldsofstudy` category (24 disciplines, median $\sim$ 433K candidates per field) as the proximity signal. We audit the false-negative rate this same-field sampling induces in Appendix B.3.

3.4 Training Configuration

Stage 2 uses sentence-transformers (Reimers and Gurevych, 2019) with MatryoshkaLoss (Kusupati et al., 2022) (output dimensions $\mathcal{K}=\{768, 512, 256, 128\}$, uniform weights) wrapping MultipleNegativesRankingLoss (Henderson et al., 2017) (InfoNCE with cosine similarity, temperature $T=0.05$). Concretely, for an in-batch set B containing the gold positive d^+, the structured hard negative d^- (Signal D), and in-batch negatives, the per-pair loss is

$$\mathcal{L}(q, d^+) = \sum_{k\in\mathcal{K}} -\log \frac{\exp\big(\mathrm{sim}_k(q, d^+)/T\big)}{\sum_{d\in B} \exp\big(\mathrm{sim}_k(q, d)/T\big)}, \tag{1}$$

where $\mathrm{sim}_k(\cdot,\cdot)$ is the cosine similarity computed on the first k embedding dimensions. Mean pooling (Nussbaum et al., 2025) and AdamW (lr 2×10^{-5}, no weight decay, gradient clip $\|g\|_2\leq1.0$, linear warmup+decay, BF16) round out the recipe. Per-device contrastive batch of 64, with 4 gradient-accumulation steps for ablation variants and 8 for SciEmbed-FULL, gives an effective optimiser batch of 256 examples (abla-

tion) or 512 examples (FULL) per parameter update; `max_seq_length`=512 for the default ablation variants (extended to 2,048 and 8,192 for the long-context variants in Section 5.4). Following the E5/GTE/Nomic recipe, each minibatch is drawn from a single signal type to prevent cross-task false negatives, with proportional sampling by available pool size (Nussbaum et al., 2025). Ablation variants train for 3 epochs at 10% warmup; SciEmbed-FULL trains for 1 epoch at 5% warmup over the full deduplicated $\sim$ 30M-pair pool. Full loss formulation (per-dimension InfoNCE with Matryoshka summation), per-signal mixing probability, and the per-pool deduplication and validation split details are in Appendix B.2.

3.5 Data Mixing

Ablation variants subsample 7M pairs (2M Signal-A triplets + 5M Signal-B context pairs) per epoch over 3 epochs ($\sim$ 1 day per variant); SciEmbed-FULL trains on the complete deduplicated A+B pool ($\sim$ 30M unique pairs after MD5 deduplication of the raw 70M-pair pool) for a single epoch. The full per-signal pool / ablation / FULL counts are in Appendix B.2 (Table 3).

4 Experimental Setup

We evaluate SciEmbed's representation quality on established scientific and long-context benchmarks.

Benchmarks. Our primary benchmark is *SciRepEval* (Singh et al., 2023): 22 tasks in four categories: classification, regression, proximity and ad-hoc search. They are evaluated via the official `allenai/scirepeval` runner with the official per-task metrics (macro-F1 / τ_b / MAP / nDCG; the 22nd task, Paper-Reviewer Matching, uses Hard/Soft P@$\{5, 10\}$ on a different scale and is reported separately; aggregation details and ordering-stability cross-checks in Appendices C.1 and C.2). As an independent leaderboard-comparable check we also evaluate the *MTEB Science Subset* (Muennighoff et al., 2023) (9 tasks) via the official `mteb` package. SciRepEval and MTEB Science are dominated by short inputs (Section 5.4) and cannot exercise the 8K backbone, so we also evaluate *LongEmbed* (Zhu et al., 2024) (6 retrieval tasks via the official upstream `mteb` runner) and a science-specific *Body-Fact Retrieval* (BFR) diagnostic probe built from S2ORC (Lo et al., 2020) full text (9,749 papers encoded as *title* + *abstract* + *body*; 1,000 held-out body-middle sentences with $\leq$ 20% trigram overlap with the abstract as queries, so the answer is in the body but not the abstract prefix; full construction in Appendix H.1). The within-recipe gap between SciEmbed-CTX variants (`max_seq_length` $\in \{512, 2048, 8192\}$) isolates the Stage-2 context-length contribution on scientific text.

Baselines. Within the single-model embedder regime (one set of weights, one embedding per document) our peers are SPECTER2 Base (Singh et al., 2023), SciNCL (Ostendorff et al., 2022), SemCSE (Brinner and Zarrieß, 2025), BGE-large-en-v1.5 (Xiao et al., 2024), BGE-M3 (Chen et al., 2024), E5-large-v2 (Wang et al., 2022), Nomic ModernBERT (Nomic AI, 2025), Nomic Embed v1.5 (Nussbaum et al., 2025), and GTE-large-v1.5 (Li et al., 2023). IBM Granite Embedding R2 (Awasthy et al., 2025) (149M, ModernBERT, 8K; trained on general enterprise retrieval data) is the matched-architecture peer used throughout to isolate the supervision-source contribution at fixed architecture. We also report SPECTER2 with its four task-specific `[CLF]`/`[QRY]`/`[RGN]`/`[PRX]` adapters (68.3 overall, in-harness re-evaluation in Appendix E.4.1) as a task-routed reference, distinct from the single-model regime; full backbone / context / parameter table in Appendix A.2.

Trained variants. We train four headline variants: **BASE** (Signal A only, 2M-pair subsample, 3 epochs), **CTX** (Signals A+B, 7M-pair subsample, 3 epochs), **NoDAPT-CTX** (CTX initialised from raw ModernBERT-base instead of the Stage-1 checkpoint), and **FULL** (A+B on the full $\sim$ 30M-pair deduplicated pool, 1 epoch). Three long-context variants **CTX-2048**, **CTX-8192**, and **NoDAPT-CTX-8192** are independent training runs of the CTX recipe with `max_seq_length` extended at Stage 2 (see Section 5.4); they are not inference-time context extensions. FULL is the recommended general-purpose artifact; CTX-8192 is the recommended long-context artifact.

Contamination audit. Because SciEmbed trains on the S2AG citation graph and SciRepEval draws from the S2AG ecosystem, we audit structural overlap: paper-level overlap against the Signal-A candidate pool is 4.2%–31.0% across tasks (full audit in Appendix D.1). No downstream label is part of the training signal in either Stage 1 (self-supervised MLM) or Stage 2 (contrastive over

text triplets); this kind of unlabeled paper-level exposure is structurally shared with SPECTER, SPECTER2, and SciNCL. We further re-evaluate the five highest-overlap tasks on a *cold-paper subset* that drops every test paper present in the pool before scoring (Appendix D.1): the cold-vs-official deltas are small ($|\Delta| \leq 3.2$) and shared in direction across SciEmbed-FULL, BGE-large, and SPECTER2, which reduces the plausibility of a training-set leakage explanation. SciEmbed has the *smallest* mean absolute delta of the three (1.20 vs. 1.38 for BGE-large and 1.33 for SPECTER2; full table in Appendix D.2).

Compute. Stage 1 MLM runs 16.1 h on 4×H200; Stage-2 training is $\sim$ 1 day on a single H100/A100 per ablation variant and $\sim$ 2 days for SciEmbed-FULL (single epoch over the $\sim$ 30M-pair pool); data generation uses SQL-level partitioning over the citation graph.

5 Results

Table 1 presents the main results on SciRepEval.

5.1 Main Results

SciEmbed-FULL reaches 66.85 ± 0.38 on SciRepEval over three seeds, surpassing BGE-large, SPECTER2 Base, and the same-backbone Nomic ModernBERT (Nomic AI, 2025). The matched-architecture peer IBM Granite Embedding R2 (Awasthy et al., 2025) reaches 67.44, which is 0.6 above SciEmbed-FULL. The pairwise model ordering is stable to within 0.15 points across three aggregation conventions (Appendix C.2). The small gap to Granite R2 hides the more informative question of *where* the two same-architecture models specialise.

Mirror-image specialisation. Granite R2 shares our backbone and recipe family but trains on general enterprise retrieval data: a mixture of publicly available retrieval pairs, IBM-internal pairs, and LLM-synthesised query–document pairs. The weights are open, but the internal and synthetic subsets are not; SciEmbed's S2AG citation-context training pairs are fully open. At fixed architecture, the two models specialise in opposite directions on SciRepEval: SciEmbed-FULL leads regression by +2.3 (28.2 vs. 25.9); Granite leads classification (+0.7), proximity (+1.7), and search (+2.2). Figure 3 renders this contrast as cohort-relative z-scores: SciEmbed is a regression outlier, while

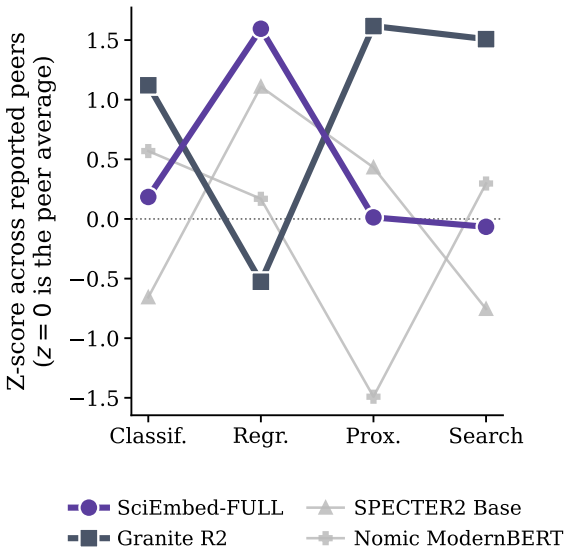


Figure 3: Per-bucket standing of SciEmbed-FULL and Granite R2 relative to SPECTER2 and Nomic ModernBERT, with each bucket normalised by its own metric. The two highlighted profiles cross between adjacent buckets: SciEmbed peaks on regression ($+1.6\sigma$, the only outlier in this bucket), while Granite peaks on proximity and search ($+1.6\sigma$ and $+1.5\sigma$) and dips on regression. No peer reproduces either shape, consistent with a supervision-source effect rather than an incidental one.

Granite peaks on proximity and search. The two profiles cross between buckets and no other peer in our cohort reproduces either shape, which is consistent with a supervision-source effect rather than recipe noise. We treat the matched-architecture pair as suggestive rather than definitive: a single comparison cannot fully rule out DAPT-corpus, batch, or data-scale differences (see Limitations).

Where the regression lead concentrates. The +2.3 regression advantage concentrates on tasks that test persistent-citation recognition (publication year, h-index, citation count); the single regression task that runs against the pattern is Tweet Mentions, a social-media metric not present in our citation-context training signal. The same +2.3 regression deficit also holds against SemCSE (Brinner and Zarrieß, 2025) (183M, LLM-summary supervision), so the advantage is not specific to general-retrieval alternatives – consistent with the lift coming from training on author-written descriptions of cited papers rather than from contrasting against a general-retrieval recipe in particular.

Task-routed reference. SPECTER2's four-LoRA-adapter system reaches 68.3 overall, with

Table 1: Main results on the official SciRepEval benchmark (Singh et al., 2023) (22 tasks via `allenai/scirepeval`). Overall is the macro-average of the four task-category averages over 21 common-metric tasks (Appendix C.1); Paper-Reviewer Matching uses P@$\{5, 10\}$ and is reported per-task in Appendix C.3. SciEmbed rows are mean $\pm$ sample-std over three seeds; per-category columns show seed means. **Bold** marks the best score in the single-model regime.

Model	Params	Classif.	Regr.	Prox.	Search	Overall
Baselines						
ModernBERT-base	149M	71.1	25.2	65.7	73.3	58.8
SPECTER2 Base	110M	74.9	27.7	81.4	81.7	66.4
SciNCL	110M	74.1	25.5	81.5	80.7	65.4
SemCSE	183M	75.3	25.9	80.2	82.4	65.9
E5-large-v2	335M	74.8	24.7	78.7	83.8	65.5
Nomic ModernBERT	149M	75.9	26.7	79.3	83.2	66.3
BGE-large-en-v1.5	335M	76.1	25.6	80.5	83.9	66.5
Granite R2	149M	**76.3**	25.9	**82.7**	**84.9**	**67.4**
SciEmbed (ours, 149M params)						
SciEmbed-BASE	149M	75.3	26.8	80.2	82.2	66.1$\pm$0.09
SciEmbed-B-only[‡]	149M	75.2	28.0	80.3	82.0	66.4$\pm$0.04
SciEmbed-CTX	149M	75.5	**28.3**	80.9	82.5	66.8$\pm$0.02
SciEmbed-NoDAPT-CTX[§]	149M	75.3	28.2	80.8	82.6	66.7$\pm$0.07
SciEmbed-FULL	149M	75.6	28.2	80.9	82.7	66.85$\pm$0.38
Task-routed reference (different deployment regime)						
SPECTER2 + 4 adapters[†]	110M $\times$ 4	77.0	31.7	81.8	82.7	68.3

[‡]Signal-B-only ablation: Stage 2 on citation-context pairs alone (no citation edges, no structured negatives), otherwise identical to SciEmbed-CTX.

[§]Stage-1-skipped variant: identical to SciEmbed-CTX (Signals A+B, 7M-pair subsample, 3 epochs) but Stage 2 is initialised from raw ModernBERT-base instead of the Stage 1 MLM checkpoint.

[†]Single-system upper bound that selects one of four task-specific LoRA adapters per downstream task; not a single-model comparison. A deployment that accepts per-task routing can exceed both single-model recipes (mostly via a specialised regression head), at the cost of dispatching a different model per task.

the gap to SciEmbed-FULL concentrated in regression (31.7 vs. 28.2) where a dedicated regression head specialises end-to-end. This is a deployment trade-off rather than a redefinition of the comparison: SciEmbed targets the single-model regime (one set of weights, one embedding per document, one index downstream), whereas deployments that can afford per-task model dispatch and are regression-dominated may prefer SPECTER2's task-routed system.

Decomposing the citation-edge contribution. Among prior citation-edge peers, SciNCL (Ostendorff et al., 2022) can be compared: graph-proximity negatives over citation edges, SciBERT backbone, 512-token cap, reaching 65.4 on SciRepEval. SciEmbed-BASE (matched citation-edge supervision but on ModernBERT-8K) reaches 66.1 ± 0.09, so the $+0.7$ from SciNCL to BASE measures the joint backbone-and-recipe upgrade. Adding citation-context queries on top of this lifts SciEmbed-CTX a further $+0.7$ that isolates the supervision upgrade itself. A separate data-scaling sweep from 3.5M to 30M pairs is monotonic and concentrated in regression (Appendix E.1).

Cross-check on MTEB Science. The same mirror-image pattern recurs on the science-relevant MTEB tasks (Muennighoff et al., 2023). SciEmbed-FULL leads ArxivClassification and MedrxivClustering outright across all eleven evaluated models (including both matched-architecture peers), while Granite R2 leads the five MS-MARCO-style retrieval tasks (SciDocsRR, SciFact, TRECCOVID, NFCorpus, BiorxivClustering). SciEmbed's lead is on the classification and clustering side; Granite's is on the retrieval side. Per-task scores are in Appendix F.

5.2 Ablation Study

We test whether supervision source drives the model specialization. Adding citation-context queries on top of citation edges (BASE $\rightarrow$ CTX) is worth $+0.7$ overall, with the lift concentrated on regression ($+1.5$). The signal exposes the model to natural-language queries that title+abstract pairs never see.

Citation context, not the edge, carries the regression lift. A Signal-B-only run – Stage 2 on citation-context pairs alone (no citation edges,

no structured negatives), otherwise identical to CTX – reaches 66.4 overall with regression 28.0 (three seeds; Table 1). Regression thus moves 26.8 → 28.0 → 28.3 across BASE (edges only) → B-only (contexts only) → CTX (both), so the regression specialisation is carried by the citation-context signal rather than the citation edge. Search runs the other way – B-only (82.0) sits just below edge-only BASE (82.2) – so the two signals are complementary rather than redundant, and the full A+B recipe is best overall.

Stage 1 MLM contributes negligibly. Removing the Stage 1 checkpoint (CTX → NoDAPT-CTX) changes the SciRepEval score by only 0.1 (66.8 → 66.7), lifts the LongEmbed aggregate by 0.5, and stays within $|\Delta| \leq 0.8$ on every BFR R@10 variant at 8K. Once Signals A+B are in the contrastive mix, DAPT's marginal contribution is essentially zero at our 8.3B-token budget.

Pool scaling saturates and alternative signals do not help. Scaling the contrastive pool from 7M to 30M pairs is monotonic on the seed-42 curve (+0.4 overall, saturating past ∼ 21M; full curve in Appendix E.1), though the three-seed FULL mean (66.85) sits only +0.05 above CTX within the seed band. Instruction-aware pairs (+0.1) and intent-conditioned training (−0.1) do not meaningfully improve over A+B (Appendix B.3). Per-category LoRA adapters (Hu et al., 2022) trade regression for the other three buckets and end at 66.6 overall, below SciEmbed-FULL (Appendix E.4).

5.3 Recipe-Space Saturation

We test whether the overall score of 66.85 reflects saturation of our runs or of citation-context supervision itself. A 28-configuration sweep over the CTX recipe varies contrastive temperature, batch and pool composition, and four alternative auxiliary signals; no variant improves on FULL by more than +0.01 (the best reaches 67.11 against the same-seed FULL baseline of 67.10; full sweep in Appendix E.3).

The flatness is not a benchmark artefact either. On half of the common-metric SciRepEval tasks, per-task variance across SciEmbed recipe variants stays under 1.0 points, while across external models on the same tasks it runs 1.5 to 5.9× larger. The benchmark has room left for moves between models; the SciEmbed recipe family does not occupy that room. We read these two probes together as evidence that, within the configurations we tested, saturation sits on the supervision-source axis, not the recipe axis.

5.4 Long-Context Retrieval

Only 3.9% of SciRepEval inputs exceed 512 tokens, so we evaluate long context separately on LongEmbed (Zhu et al., 2024) and on a science-specific Body-Fact Retrieval (BFR) diagnostic probe (Section 4). The same mirror-image pattern from Section 5.1 repeats: Granite R2's general-retrieval supervision wins absolute scores on both, while within the SciEmbed recipe a Stage-2 context-length scan (512 → 2K → 8K) is monotonically positive on both benchmarks. Two BFR robustness variants preserve this monotonicity after restricting candidates to same-field neighbours or rewriting queries with Qwen3-32B, ruling out coarse topic separability and exact-string matching as alternative explanations. We treat BFR as a diagnostic rather than a benchmark. Per-length NeedleRetrieval, Matryoshka truncation, per-task LongEmbed curves, BM25 lexical controls, and the full BFR robustness tables are in Appendices E.2, G.2, G.4, H.2.

5.5 Efficiency

In fp16 on a single H100, SciEmbed processes 5,363 docs/sec at bs=128, exceeding BGE-large (4,875 docs/sec) by 10% while using 2.6× less peak GPU memory (445 MB vs. 1,168 MB). The same advantage holds on a workstation-class GPU (2,004 vs. 1,133 docs/sec, 1.77×), so the efficiency is not datacenter-specific. SPECTER2 reaches a higher absolute throughput but is restricted to a 512-token input window. The same-size Granite R2 matches SciEmbed's throughput to within 1% (Appendix I.1), so the efficiency gap is with the 335M-class models rather than between the two 149M peers. The full throughput curve, the small-batch latency table, and the UMAP visualisation are in Appendices I.1 and I.2; per-task SciRepEval scores in Appendix C.3.

6 Conclusion

We present SciEmbed, a scientific document embedder that uses the citation context sentence, the description a citing author writes about a cited work, as its primary contrastive supervision signal. On SciRepEval it reaches 66.85, surpassing prior open scientific embedders such as SPECTER2 and SciNCL and coming within 0.6 points of the same-

size Granite R2, which is trained on proprietary general-retrieval data.

A matched-architecture comparison against Granite R2 is consistent with a supervision-source effect: with backbone, parameter count, and recipe family held fixed, the two models specialise in opposite directions, SciEmbed on the impact-related regression tasks and Granite on classification, proximity, and search. Our recipe and pretraining ablations leave this picture unchanged, so the specialisation is best explained by the supervision source, not by architecture, scale, or training recipe, though the closed composition of Granite's training data leaves a residual data confound we cannot fully exclude (see Limitations).

This positions the supervision source as a primary design choice for scientific document embeddings, alongside architecture and training recipe. Citation-context and general-retrieval supervision act as complementary signals, each strongest on the tasks it targets: fine-tuning Granite R2 on our citation contexts adds the regression specialisation to a strong retrieval backbone for the best overall of the three variants (67.6; Appendix E.5), confirming the two compose rather than trade off. We release SciEmbed, its training pipeline, and evaluation harness so the community can build on the citation-context axis and combine it with others.

Limitations

Data and language. SciEmbed is trained exclusively on English-language papers and inherits coverage biases from S2AG, which underrepresents certain fields, regions, and venues. Deployments that rely on the impact-related regression scores should pair them with field-normalized evaluation and fairness checks across underrepresented fields, languages, and regions.

General-retrieval and long-context gap. Citation-context supervision does not close the gap to general-purpose embedders on retrieval-style tasks. SciEmbed trails BGE-large on SciRepEval search (82.7 vs. 83.9), scores 49.3 on the full 41-task MTEB(eng,v2) suite against Nomic Embed v1.5's 63.4, and trails Granite R2, BGE-M3, and GTE-large-v1.5 on LongEmbed.

Matched-architecture comparison. The Granite R2 baseline shares our backbone, parameter count, and recipe family but trains on a mixture that includes IBM-internal and LLM-synthesised query-document pairs whose composition and scale are not public. We can therefore isolate the supervision *style* (citation-context vs. general retrieval) only conditional on Granite's training-data scale and quality being broadly comparable. We mitigate this by triangulating against multiple matched-architecture peers (Nomic ModernBERT, SemCSE) on both SciRepEval and MTEB Science, and by the recipe-space ceiling probe (Appendix E.3) that shows no SciEmbed configuration we tested exceeds the headline, but the comparison cannot fully rule out a data-scale or data-quality confound from Granite's closed corpora.

NoDAPT finding is budget-conditional. The negligible Stage-1 contribution holds at our 8.3B-token DAPT budget and single-phase cosine schedule. We did not vary the Stage-1 corpus, the MLM objective, or the architectural scale, and a longer or differently scheduled DAPT could still contribute, particularly where token-level vocabulary adaptation is the dominant signal.

Construction caveats. Long-context experiments use a single training seed; the within-recipe monotonicity claim is replicated across four probes (LongEmbed, which includes NeedleRetrieval, and three BFR variants). BFR itself is a self-constructed probe and is framed as a diagnostic rather than a benchmark. Signal B treats every S2AG citation context as a clean positive without filtering perfunctory or self-citations (Wu et al., 2026): a frontier-LLM audit of 2,000 pairs (Appendix B.3) finds the cited title+abstract recovers the referenced aspect in 62.5% of pairs and misses it in 37.5%, so the +0.7 overall lift is an aggregate-signal result that tolerates this noise (the Signal-B-only ablation in Section 5.2 isolates it end-to-end). Selecting higher-quality contexts – a substantive-context or citation-quality filter over Signal B, together with self-citation removal – is a natural next step we leave to future work. Symmetrically, Signal D draws same-field negatives at the `s2fieldsofstudy` granularity, rejecting only the anchor and the gold positive, so a paper genuinely related to the anchor can be sampled as a negative. An audit of 1M training negatives (Appendix B.3) finds 4.7% of same-field negatives are directly cited by or cite the anchor and 7.3% share any citation or bibliographic-coupling link, against a ∼0% random-field floor, so this false-negative contamination is small and, like the positive side, leaves Signal D an aggregate-signal result. Text-

level MD5 deduplication of training pairs collapses identical citing phrasings of the same target paper into a single example, which flattens the underlying citation-frequency distribution.

Recipe and architectural scope. All variants use mean pooling and Matryoshka loss with uniform weights across $\{768, 512, 256, 128\}$; attention-weighted and CLS pooling, late-interaction architectures such as ColBERTv2 (Santhanam et al., 2022), dimension-specific temperatures, and capacity-aware Matryoshka weights are not ablated. The recipe-space saturation analysis is bounded to a single backbone, parameter count, and recipe family; we do not test substantially larger backbones, teacher distillation, or cross-encoder heads. Signal B also uses the cited title+abstract as the positive; aligning the citation context to the most relevant passage of the cited paper, in the style of CL-SciSumm cited-text-span annotation (Jaidka et al., 2016), is a promising richer-supervision direction our released pipeline supports.

Ethics Statement

All inputs to SciEmbed are publicly available academic metadata and full-text papers under open-access licenses (S2ORC, PMC Open Access Subset, arXiv, peS2o); no personally identifiable information is collected. Compute and infrastructure are reported in Section 4.

Coverage bias and downstream use. S2AG over-represents English-language, North American and European, biomedical, and STEM publications, and the citation-context (Signal B) signal is denser in biomedical and computer-science papers than in humanities-adjacent fields. These biases propagate into the trained model and any downstream system built on it. We do not recommend SciEmbed as the sole signal in high-stakes settings such as reviewer assignment, hiring, or funding decisions without additional fairness-aware controls.

Misuse and release. SciEmbed could be misused to construct citation cartels or amplify the coverage biases above. We release weights, training pipeline, evaluation runner, and the project-level `uv.lock` under the MIT License with a model card documenting these limitations.

Acknowledgments

The authors gratefully acknowledge the computing resources provided by the Center for Information Services and High Performance Computing (ZIH) at TU Dresden, in particular the Capella and Alpha Centauri GPU clusters, on which SciEmbed was trained and evaluated.

Use of AI assistants. We disclose two specific uses of AI assistance. (i) *Language polish:* an AI writing assistant was used for sentence-level grammar and style touch-ups on portions of the manuscript text. (ii) *Coding utility:* an AI coding assistant was used to draft scripts (e.g., data inspection, plotting, log parsing). All claims, experimental design choices, training and evaluation runs, ablations, error analysis, and the discussion and interpretation of results are the authors' own; no section, paragraph, table, or claim in the manuscript was entirely generated by an AI system, and no AI was used to select related work or to generate citations.

References

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, P Vignesh, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. 2025. Granite embedding R2 models. *arXiv preprint*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics.

Marc Felix Brinner and Sina Zarrieß. 2025. SemCSE: Semantic contrastive sentence embeddings using LLM-generated summaries for scientific abstracts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32704–32719. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3586–3596. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282. Association for Computational Linguistics.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the CL-SciSumm 2016 shared task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 93–102.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, and 29 others. 2023. The Semantic Scholar open data platform. *arXiv preprint*.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems 35*, volume 35, pages 30233–30249. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations*.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025b. Clinical ModernBERT: An efficient and long context encoder for biomedical text. *arXiv preprint*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint*.

LightOn AI. 2025. modernbert-embed-large: A general-purpose embedding model based on ModernBERT-large. `https://huggingface.co/lightonai/modernbert-embed-large`.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037. Association for Computational Linguistics.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470. Association for Computational Linguistics.

NeuML. 2025. bioclinical-modernbert-base-embeddings: Sentence-transformers fine-tune of BioClinical ModernBERT on PubMed pairs. `https://huggingface.co/NeuML/bioclinical-modernbert-base-embeddings`.

Nomic AI. 2025. modernbert-embed-base: General-purpose embedding model based on ModernBERT. `https://huggingface.co/nomic-ai/modernbert-embed-base`.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. Nomic Embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 213–222. ACM.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566. Association for Computational Linguistics.

Shruti Singh and Mayank Singh. 2024. CoSAEmb: Contrastive section-aware aspect embeddings for scientific articles. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 283–292. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788. Association for Computational Linguistics.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. BioClinical ModernBERT: A state-of-the-art long-context encoder for biomedical and clinical NLP. *arXiv preprint*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2025. Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, pages 123–129. Springer Nature Switzerland.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

The Mosaic ML Team. 2021. Composer. `https://github.com/mosaicml/composer/`.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-worthiness detection for improved scientific document understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1796–1807. Association for Computational Linguistics.

Huidong Wu, Haojia Xiang, Jingtong Gao, Xiangyu Zhao, Dengsheng Wu, and Jianping Li. 2026. Detecting miscitation on the scholarly web through LLM-augmented text-rich graph learning. In *Proceedings of the ACM Web Conference 2026*, pages 4208–4219. ACM.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649. Association for Computing Machinery.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint*.

Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023. Pre-training multi-task contrastive learning models for scientific literature understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12259–12275. Association for Computational Linguistics.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816. Association for Computational Linguistics.

A Reproducibility

A.1 Software and dependencies

The released `uv.lock` pins the full dependency graph. Stage 1 runs MosaicML Composer (The Mosaic ML Team, 2021) 0.32.1 on PyTorch 2.7.0 in BF16, with `transformers` 4.57.6, Flash Attention 2.8.3, and an 8.3B-token `mosaicml-streaming` 0.13.0 MDS shard. Stage 2 uses `sentence-transformers` 5.3.0 on a single A100 or H100 in BF16. Data construction uses DuckDB 1.5.0 for SQL-level partitioning of the citation graph, PyArrow 23.0.1, and `datasets` 4.8.2 for in-RAM signal mixing. Evaluation uses the official `allenai/scirepeval` protocol on the 22-task benchmark and `mteb` 2.12.15 for MTEB and LongEmbed. Python $\geq$ 3.12 throughout.

A.2 Open embedding-model landscape

Table 2 lists a selection of open embedding models with parameter count, native context length, specialization and backbone.

B Training construction

B.1 Stage 1: domain-adaptive MLM

Stage 1 MLM starts with 13.9M scientific full-text papers from four sources: PubMed Central Open Access Subset (PMC-OAS; 4.6M), S2ORC (Lo et al., 2020) (7.4M), arXiv (748K), and peS2o (Soldaini et al., 2024) (1.1M), filtered to 1,000–500,000 characters of body text, yielding 13.4M papers.

Text cleaning. We remove LaTeX artefacts (while trying to preserve meaningful mathematical notation), strip XML markup from PMC documents, remove reference sections and publisher boilerplate, normalise Unicode, and collapse whitespace.

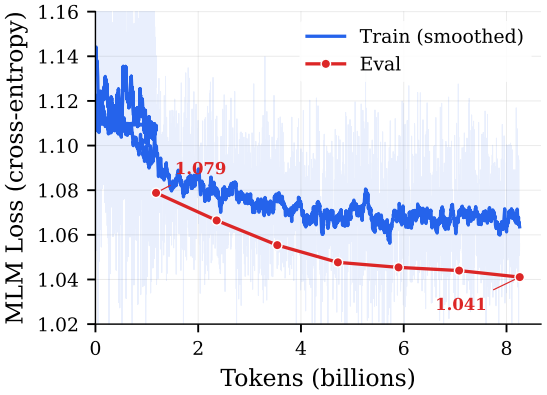


Figure 4: Stage 1 MLM training and validation loss over 8.3B tokens on 4×H200 GPUs. Eval cross-entropy decreases from 1.079 to 1.041 (−3.5%).

Tokenisation and chunking. The ModernBERT BPE tokenizer (50,368-token vocabulary, byte-level pre-tokenization) splits each document into 8,192-token sequences with 128-token stride overlap. We sample 8–10B tokens from the corpus.

Training recipe. Following the ModernBERT architecture (Sounack et al., 2025; Warner et al., 2025) we use a 30% masking rate. The MLM objective is the standard masked-token negative log-likelihood,

$$\mathcal{L}_{\text{MLM}} = -\sum_{t \in \mathcal{M}(x)} \log p_\theta\left(x_t \mid x_{\setminus \mathcal{M}(x)}\right), \quad (2)$$

where $\mathcal{M}(x) \subset \{1, \ldots, |x|\}$ is the random 30%-masking subset and $x_{\setminus \mathcal{M}(x)}$ is the unmasked context. We optimise $\mathcal{L}_{\text{MLM}}$ with AdamW (β_1=0.9, β_2=0.95, weight decay 0.01), learning rate 2×10^{-5}, 2,000 warmup steps, cosine annealing, and BF16 mixed precision. Training runs on 4 NVIDIA H200 GPUs via MosaicML Composer (The Mosaic ML Team, 2021) for 16.1 h (8.3B tokens at 143.6K tok/s steady-state). Training and validation loss can be seen in Figure 4.

B.2 Stage 2: contrastive fine-tuning

Stage 2 contrastive fine-tuning uses sentence-transformers (Reimers and Gurevych, 2019) with MatryoshkaLoss (Kusupati et al., 2022) wrapping MultipleNegativesRankingLoss (Henderson et al., 2017). For each batch $\mathcal{B}$ of anchor/positive pairs (q_i, d_i^+) (optionally accompanied by per-anchor hard negatives $\mathcal{N}_i$ from Signals A and D), MNRL optimises an InfoNCE objective with cosine simi-

Table 2: Open embedding models, early 2026. "Sci" = trained on scientific signals; "Ctx" = native context length.

Model	Params	Ctx	Sci	Backbone
SPECTER2 (Singh et al., 2023)	110M	512	✓	SciBERT
SciNCL (Ostendorff et al., 2022)	110M	512	✓	SciBERT
SciMult (Zhang et al., 2023)	110M	512	✓	SciBERT
SemCSE (Brinner and Zarrieß, 2025)	183M	512	✓	SciDeBERTa
BGE-large (Xiao et al., 2024)	335M	512	–	BERT-large
E5-large (Wang et al., 2022)	335M	512	–	BERT-large
GTE-large-v1.5 (Li et al., 2023)	434M	8192	–	BERT+RoPE+GLU
Nomic Embed v1 (Nussbaum et al., 2025)	137M	8192	–	long-context BERT
Jina-v3 (Sturua et al., 2025)	570M	8192	–	jina-XLM-R
BGE-M3 (Chen et al., 2024)	568M	8192	–	XLM-R-large
NV-Embed-v2 (Lee et al., 2025a)	8B	32K	–	Mistral-7B
Qwen3-Emb-8B (Zhang et al., 2025)	8B	32K	–	Qwen3
nomic-modernbert-embed (Nomic AI, 2025)	149M	8192	–	ModernBERT
lighton-modernbert-embed-l (LightOn AI, 2025)	395M	8192	–	ModernBERT
Granite R2 (Awasthy et al., 2025)	149M	8192	–	ModernBERT
BioClin. ModernBERT (Sounack et al., 2025)	149M	8192	(biomed MLM)	ModernBERT
NeuML bioclin-mb-emb (NeuML, 2025)	149M	8192	(biomed sim.)	ModernBERT
SciEmbed (ours)	**149M**	**8192**	✓(citation graph)	ModernBERT

larity $s(q, d) = \cos(f_\theta(q), f_\theta(d))$:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{|\mathcal{B}|} \sum_i \log \frac{e^{s(q_i, d_i^+)/T}}{\sum_{d \in \mathcal{C}_i} e^{s(q_i, d)/T}}, \quad (3)$$

where $\mathcal{C}_i = \{d_j^+\}_j \cup \mathcal{N}_i$ is the per-anchor candidate set: every in-batch positive d_j^+ (with d_j^+ for $j \neq i$ acting as an in-batch negative for q_i) plus any explicit hard negatives $\mathcal{N}_i$ ($\mathcal{N}_i = \varnothing$ for pair-only signals such as Signal B alone). $T = 1/20 = 0.05$ is the softmax temperature (sentence-transformers `scale`=20); with the per-device batch of 64, each anchor sees at least 63 in-batch negatives plus its hard negative when present. The effective optimiser batch (256 for ablation variants, 512 for SciEmbed-FULL) is produced by gradient accumulation, so in-batch negatives still come from the 64-example forward step.

Let $\mathcal{D} = \{128, 256, 512, 768\}$ denote the four Matryoshka truncation dimensions. The wrapper applies Eq. 3 at each $d \in \mathcal{D}$ with uniform weights $w_d = 1$:

$$\mathcal{L} = \sum_{d \in \mathcal{D}} w_d \, \mathcal{L}_{\text{InfoNCE}}^{(d)}. \quad (4)$$

The default embedding dimension at inference is 768.

Following the E5/GTE/Nomic recipe, each minibatch is drawn from a single signal type to prevent cross-task false negatives. With sentence-transformers' `MultiDatasetBatchSamplers.PROPORTIONAL` sampler, signal $s \in \{A, B\}$ is sampled with probability $p(s) \propto |\mathcal{S}_s|$, where $\mathcal{S}_A$ and $\mathcal{S}_B$ are the per-signal pair pools (ablation regime: $|\mathcal{S}_A|$∼2M, $|\mathcal{S}_B|$∼5M; we use $\mathcal{S}$ here to avoid overloading $\mathcal{D}$). A 5% held-out validation split monitors training with early stopping (patience 3–5).

The $\sim 57\%$ reduction from the 70M raw pool to $\sim$ 30M unique pairs reflects three properties of the S2AG graph. (i) The raw `s2ag.citations` table is itself not edge-unique: $\sim 9\%$ of rows are duplicate $(citing, cited)$ pairs (2.92B rows vs. 2.65B distinct edges). (ii) S2AG contains multiple corpus identifiers for the same underlying paper (preprint, journal, version-of-record), so canonical $(\text{title}, \text{abstract})$ texts collide across distinct `corpusids`. (iii) Signal B is dominant: the same citation context sentence frequently appears in multiple citing papers pointing to the same target (boilerplate phrases such as "Following $X \ldots$" for canonical references). MD5 on the formatted anchor/positive text collapses all three to a single training pair, so any two pairs that would present identical training signal to the encoder are merged.

B.3 Construction details for ablation and probe signals

This subsection covers two groups of signals outside the main A+B recipe: the recipe-side ablation signals (C, D), and the four data-signal sources used in the ceiling probe (Section 5.3). The Signal A+B construction itself is illustrated in Figure 1.

Table 3: Training signal summary for Stage 2 contrastive fine-tuning. "Pool" is the total generated pairs; "Ablation" is the subsample drawn per epoch ($\times 3$ epochs); "Full" is unique pairs after MD5 deduplication ($\times 1$ epoch). Signal D is applied to A–B rather than contributing independent pairs. Signal C and instruction pairs were explored but did not improve on A+B (Section 5.2).

Signal	Source	Pool	Ablation	Full	Prefix
A: Citation edges	S2AG citations	50M	2M	∼25M	`search_*`
B: Context queries	S2AG contexts	20M	5M	∼5M	`search_*`
D: Hard negatives	S2AG + fields	–	–	–	(applied to A–B)
Explored signals (not in the headline recipe)					
C: Intent-conditioned	S2AG intents	15M	–		`cite_*`
Instruction pairs	OpenAlex	10M	–		`classify:/search_*`
Total (A+B)		**70M**	**7M**	**∼30M***	

*After MD5 deduplication of the 70M raw pool to ∼30M unique pairs.

Signal C (intent-conditioned, discarded). S2AG provides citation intent labels (`background`, `method`, `result`) inferred from context. We construct intent-prefixed query/document pairs of the form `cite_method:` *{context}* $\rightarrow$ *{cited title+abstract}*. The raw distribution is heavily skewed (approximately background 80%, method 18%, result 4%; individual class fractions are independently rounded and need not sum exactly to 100%, since S2AG citations may carry more than one intent label). We oversample method $3\times$ and result $10\times$ subject to a global cap of 15M pairs, yielding approximately 6.9M background, 4.6M method, and 3.4M result pairs. Intent-specific hard negatives draw from same-field papers cited under the *other* intents.

Signal D (structured hard negatives). We add an explicit structured hard negative to the standard in-batch negatives, sampled from three strata drawn from a pre-materialised pool of S2AG papers and field-of-study labels: **same-field non-cited (50%)** forces the model to distinguish topical proximity from intellectual debt; **forward-neighbour papers (30%)** are papers cited by the cited (positive) paper, drawn from a forward citation index over Signal-A candidates; **random (20%)** provide easy negatives for calibration. Signal D is applied to Signals A and B rather than contributing standalone pairs, and is integrated into the headline SciEmbed-FULL recipe (Table 1).

False-negative audit of same-field negatives. Because the same-field stratum samples uniformly within a field (median ∼433K candidates) and rejects only the anchor and the gold positive, a paper genuinely related to the anchor can be drawn as a negative. We bound how often this happens by scoring 1M sampled training triplets against the citation graph, using two relatedness criteria that are independent of the positive citation edge: a direct citation between anchor and negative in either direction (other than the gold positive), and bibliographic coupling of at least three shared references. Among same-field negatives, 4.7% are directly cited by or cite the anchor and 7.3% are related by either criterion, against a ∼0% random cross-field floor and 100% direct citation for gold positives (Table 4); roughly 93% carry no citation or bibliographic-coupling link to the anchor. The false-negative contamination is therefore small and bounded, and Signal D is, like the positive side, an aggregate-signal mechanism rather than a guarantee that every sampled negative is unrelated.

Table 4: False-negative audit of Signal-D negatives, scored on 1M sampled training triplets against the citation graph (criteria independent of the positive edge). *Cited*: a direct citation in either direction other than the gold positive; *BC≥3*: bibliographic coupling of ≥ 3 shared references; *Any*: either criterion. Rates in %. Gold positives and random cross-field pairs are calibration references; bibliographic coupling is a coarse proxy, as even gold positives reach only 49.9 at BC≥3.

Pair population	Cited	BC≥3	Any
Same-field negatives	4.7	5.5	7.3
All negatives (mix)	3.9	4.3	6.0
Gold positives (ref.)	100.0	49.9	100.0
Random cross-field (floor)	0.0	0.0	0.0

Context–abstract alignment audit. Signal B assumes the cited title+abstract is a usable positive for the citation-context query. We audit this assumption via a frontier LLM judge (Claude Opus 4.8). We label 2,000 randomly sampled Signal-B pairs on two axes and assess whether the specific aspect

Table 5: Context–abstract alignment audit on 2,000 randomly sampled Signal-B pairs, judged by a frontier LLM. *Aligned* = FULL (aspect explicit in title/abstract) + PARTIAL (recoverable from the abstract topic); *Subst.* = the context describes the cited work rather than listing it in passing. Rates in % over parsed judgments.

Judge	FULL	PARTIAL	MISMATCH	Subst.
Opus 4.8 (primary)	20.4	42.1	37.5	83.9
DeepSeek-V4 (check)	22.3	28.6	49.1	73.3

the citation attributes to the cited paper is recoverable from its title+abstract (FULL: stated explicitly; PARTIAL: clear from the abstract's topic; MISMATCH: not recoverable), and whether the context substantively describes the cited work. The aspect is recoverable in 62.5% of pairs (20.4% FULL + 42.1% PARTIAL) and a MISMATCH in 37.5%; 83.9% of contexts are substantive (Table 5). A second judge (DeepSeek-V4) reproduces the majority-aligned-but-noisy picture (50.9% aligned, agreeing with Opus on 79.9% of items), and author spot-checks concur, so the finding is not an artifact of a single judge. The signal is thus noisy but works through the noise: InfoNCE with in-batch negatives tolerates imperfect positives, and the Signal-B-only ablation (Section 5.2) shows the context signal alone still carries the regression specialisation. Selecting for the recoverable majority (a substantive-context filter over Signal B) is a natural improvement we leave to future work.

Ceiling-probe data signals. Four data sources were introduced specifically for the Section 5.3 ceiling probe. Each was evaluated both as a pure replacement for Signal A+B and mixed into A+B. None exceeded the FULL baseline; the per-configuration scores are in Table 12.

OpenAlex related-works pairs. A paper-paper topical-similarity edge set drawn from OpenAlex's `works_related_works` relation. We extracted 30M filtered (anchor, related) pairs in the same parquet schema as Signal A+B, with English-language and abstract-length filters ($50 \leq |\text{abstract}| \leq 5000$ characters). At 30M pairs they ranked lowest of the four probe signals (pure 63.25, mixed 65.08).

Co-citation pairs. The classical scientometric "co-citation" similarity signal: two papers are co-citation siblings when they share a citing parent. We extracted 1.19M pairs from a self-join of `s2ag.citations` on the citing paper, restricted to a 5M citing-paper sample with at most 10 outgoing citations per parent. This is structurally the supervision underlying SciDocs CoCite, CoView, and CoRead, where we trail Granite R2 by -1.1 on average. Empirically: pure 65.37, mixed 66.53.

TLDR↔abstract pairs. The model-generated TLDR of a paper (from S2AG's `tldrs` table, ∼70M paper-TLDR pairs) is treated as a query for the paper's full abstract. We extracted 15.55M filtered pairs subject to the same length and English filters as Signal A+B. The bet was that an asymmetric short-query/long-document structure is closer to MS-MARCO retrieval than the symmetric citation-context structure of A+B and might lift the search bucket. The search score on the pure variant is 80.19, -2.50 below the NoDAPT-CTX recipe base (82.69) on which the probe is built. Empirically: pure 63.60, mixed 65.46.

Intent-filtered improvement citations. A drug-discovery-style triplet: the anchor is a cited (prior) paper, the positive is the citing (follow-up) paper, restricted to citations with `is_influential = TRUE` in S2AG (approximately 12% of citations). We drew 2.0M pairs by modular sampling on `citationid`. Empirically: pure 65.38, mixed 66.63, the closest call in the entire probe.

C Evaluation protocol

C.1 SciRepEval task aggregation

Table 6 maps each SciRepEval task to its primary metric and the bucket used in our four-category macro-average (the "Overall" column of Table 1). We follow the per-task metrics of Singh et al. (2023) with two deviations from their format-based grouping: RELISH (nDCG) is scored alongside the other nDCG-based ad-hoc retrieval tasks rather than the proximity tasks, and Paper-Reviewer Matching (PRM) is excluded from the four-category macro because its Hard/Soft P@k scale would distort a category mean. PRM is still reported per metric in Table 8, and the cross-check in §C.2 shows that the headline ordering is robust to both choices.

C.2 Aggregation cross-check

To check that the headline ordering is not an artefact of our four-category macro, Table 7 recomputes the overall score under three conventions: (A) our four-category macro over the 21 common-metric tasks; (B) the per-task mean over all 22 tasks, with PRM contributing the mean of its four P@$\{5, 10\}$ sub-metrics; (C) a 5-format macro treat-

Table 6: SciRepEval task to category mapping. "Aggr." is whether the task contributes to the four-category macro-average reported in Table 1.

Task	Metric	Bucket	Aggr.
MeSH Descriptors	F1-macro	Classification	yes
FoS	F1-macro	Classification	yes
Biomimicry	F1-binary	Classification	yes
DRSM	F1-macro	Classification	yes
SciDocs MAG	F1-macro	Classification	yes
SciDocs MeSH	F1	Classification	yes
Citation Count	Kendall τ_b	Regression	yes
Publication Year	Kendall τ_b	Regression	yes
h-Index	Kendall τ_b	Regression	yes
Tweet Mentions	Kendall τ_b	Regression	yes
Peer-Review Score	Kendall τ_b	Regression	yes
Same Author	MAP	Proximity	yes
Highly Infl. Cite.	MAP	Proximity	yes
SciDocs Cite	MAP	Proximity	yes
SciDocs Co-Cite	MAP	Proximity	yes
SciDocs Co-View	MAP	Proximity	yes
SciDocs Co-Read	MAP	Proximity	yes
Search	nDCG	Search	yes
TREC-CoVID	nDCG	Search	yes
NFCorpus	nDCG	Search	yes
RELISH[†]	nDCG	Search	yes
Paper-Reviewer Match.	H/S P@5,@10	—	no

[†]RELISH is officially a proximity task in Singh et al. (2023) but uses nDCG; regrouping it with the nDCG tasks shifts the four-category overall by 0.3–0.5 points and preserves the top-4 ordering.

Table 7: SciRepEval overall score under three aggregation conventions. (A) our 4-cat macro (21 tasks). (B) 22-task mean with PRM folded in. (C) 5-format macro with PRM as its own format. The model ordering is identical under all three; pairwise gaps are stable to within 0.15 points. SciEmbed-FULL averages over seeds 123, 456, and 789.

Model	A	B	C
SPECTER2 + 4 adapters	68.30	67.62	63.72
Granite Embedding R2	67.44	66.77	63.16
SciEmbed-FULL (3 seeds)	66.85	66.21	62.65
BGE-large-en-v1.5	66.52	65.85	62.36
SPECTER2 Base	66.43	65.84	62.23
Nomic ModernBERT	66.27	65.59	62.15

ing PRM as its own format. Convention B mirrors the per-task average reported by Singh et al. (2023), who additionally apply In-Train / Out-of-Train weighting and the SciDocs conventions of Cohan et al. (2020).

The absolute level shifts as PRM is folded in (its P@$\{5, 10\}$ scale runs lower than F1/τ/MAP/nDCG), but pairwise comparisons are preserved: SciEmbed-FULL trails Granite R2 by -0.59, -0.56, -0.51 under A, B, C and

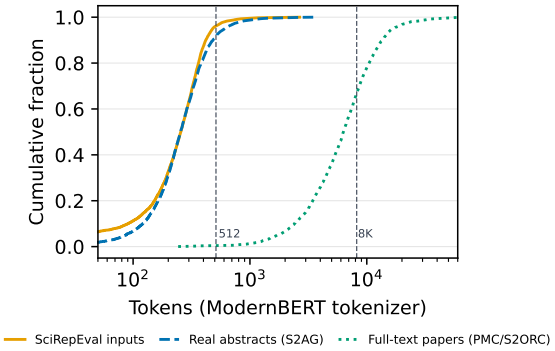


Figure 5: Cumulative length distribution of SciRepEval inputs vs. natural scientific text. SciRepEval inputs nearly match raw abstracts in the bulk but have a thinner right tail (3.9% vs. 8.2% over 512 tokens), consistent with construction-time selection for SciBERT-class encoders. Full-text papers live almost entirely beyond 512 tokens, with 33% requiring more than 8,192.

exceeds BGE-large by $+0.33$, $+0.36$, $+0.29$. The remaining gap to the 70.9 Adapters figure of Singh et al. (2023) (and the 71.2 they obtain by stacking MTL CTRL on top of Adapters) is largely from the SciDocs and In-Train / Out-of-Train conventions we do not replicate; the per-category directions of effect are consistent across all three aggregations.

C.3 Per-task SciRepEval scores

Table 8 reports per-task scores for SciEmbed-BASE, -CTX, -NoDAPT-CTX, and -FULL alongside the strongest baselines on all 22 SciRepEval test tasks. SciDocs MeSH (F1) is grouped with classification; RELISH (nDCG) with ad-hoc search; Paper-Reviewer Matching is shown separately because its primary metrics are H/S P@5,10. SciEmbed columns are representative single-seed runs from the multi-seed cohort; per-category multi-seed means appear in Table 1.

C.4 Input length distribution

Our pipeline uses long context asymmetrically: Stage 1 MLM processes the 13.4M filtered full-text papers at the full 8,192-token window from ModernBERT, while Stage 2 contrastive fine-tuning runs on title+abstract and citation-context inputs at `max_seq_length=512`. We tokenize a stratified sample of inputs from five SciRepEval tasks (FoS, Biomimicry, Citation Count, MeSH, DRSM; 1,500 each) with the ModernBERT tokenizer, and compare against (i) 5,000 raw S2AG abstracts and (ii) 3,000 full-text papers from PubMed Central and S2ORC.

Table 8: Per-task SciRepEval scores (official evaluation). Best in **bold**, second-best underlined. SciEmbed columns are representative single-seed runs (the multi-seed mean is in Table 1). "GR2" is Granite R2; "NoDAPT" is SciEmbed-NoDAPT-CTX.

Task		SPEC2	BGE-l	Nomic	GR2	BASE	CTX	NoDAPT	**FULL**
Classification (F1)									
MeSH	F1	86.8	87.9	87.4	**88.7**	86.9	86.8	86.9	87.3
FoS	F1	42.2	**48.1**	45.6	46.9	45.0	46.5	46.3	46.4
Biomim.	F1	74.2	72.5	73.1	73.0	73.4	73.8	73.3	**74.6**
DRSM	F1	76.5	75.2	**77.9**	75.3	74.5	75.8	75.3	75.7
MAG	F1	81.0	82.6	82.1	**82.7**	81.1	81.8	81.1	81.7
SciDocs MeSH	F1	89.0	90.1	89.2	**91.2**	89.3	88.9	88.6	89.7
Regression (τ_b)									
Cite cnt	τ_b	**38.4**	36.4	37.4	35.5	35.6	37.3	37.3	38.0
Pub year	τ_b	35.6	32.0	33.1	33.9	34.0	39.2	39.2	**39.6**
hIndex	τ_b	15.8	16.0	15.8	14.0	16.9	18.2	18.0	**18.6**
Tweets	τ_b	27.7	26.1	27.6	**28.5**	27.4	27.6	27.7	28.4
Review	τ_b	20.8	17.3	19.3	17.6	**21.8**	18.2	18.5	20.1
Proximity (MAP)									
Cite	MAP	92.0	88.6	87.1	**93.2**	89.0	89.4	89.3	89.7
Co-cite	MAP	91.7	89.8	88.4	**92.3**	89.8	90.4	90.4	90.7
Co-view	MAP	85.5	84.6	83.8	**85.8**	83.9	84.8	84.5	84.6
Co-read	MAP	87.2	86.0	85.5	**87.8**	86.6	87.3	87.2	87.5
Hi-infl.	MAP	45.0	47.1	45.8	**48.1**	45.0	45.9	45.9	45.9
Author	MAP	87.0	87.0	85.4	**88.7**	86.9	87.7	87.5	87.7
Ad-hoc Search (nDCG)									
Search	nDCG	73.8	74.5	74.2	**75.5**	73.1	73.5	73.6	73.8
TREC-C	nDCG	89.5	91.0	92.1	**92.9**	91.5	91.2	91.2	91.5
NFC	nDCG	72.0	**78.0**	74.6	77.7	73.1	74.2	74.5	74.9
RELISH	nDCG	91.6	92.3	91.9	**93.5**	91.6	91.5	91.5	91.7
Recommendation: Paper-Reviewer Matching[†]									
PRM	H-P@5	34.0	34.2	**34.8**	34.4	34.4	34.6	34.8	34.2
PRM	H-P@10	25.4	25.6	25.4	**25.8**	25.8	25.7	25.6	25.8
PRM	S-P@5	66.7	**67.7**	67.1	67.7	66.7	67.7	67.7	67.1
PRM	S-P@10	55.5	55.5	55.4	**56.3**	55.9	55.8	55.8	56.0

[†]Hard (H) counts a top-k candidate as correct only if the reviewer-paper match is in the highest ground-truth bucket of Singh et al. (2023); Soft (S) includes all positive buckets. Not aggregated into the four-category macro of Table 1.

Figure 5 shows the resulting cumulative distributions. Only 3.9% of SciRepEval inputs exceed 512 tokens (median 256, 95th percentile 480, max 2,664); raw S2AG abstracts contain roughly twice as many over-512 inputs (8.2%, 95th percentile 610). The simplest explanation is that SciRepEval was assembled around 512-token encoders. Real scientific papers are dramatically longer than either (median 6,314 tokens, 33% above 8,192), so SciRepEval cannot reward an 8K-context encoder for processing long inputs, which motivates LongEmbed and BFR as complementary axes.

D Contamination audit

We analyze and quantify structural overlap between SciRepEval test tasks and the SciEmbed training dataset.

D.1 Per-task overlap

Table 9 reports overlap between every SciRepEval test task and two upper-bound training universes (cf. §4), computed against the 2026-03-01 s2ag snapshot using (`s2ag.papers`, `s2ag.citations` and `fulltext.papers`). Non-integer SciDocs identifiers are resolved to S2 corpus ids via the official `scidocs_view_cite_read` and `scidocs_mag_mesh` meta configs; ad-hoc search corpora (`search`, `nfcorpus`, `trec_covid`) use benchmark-specific identifiers not directly mappable to S2AG.

D.2 Cold-paper subset evaluation

We re-evaluate the five highest-overlap classification and regression tasks (Biomimicry, MeSH Descriptors, SciDocs MeSH, Citation Count, Publication Year) on a *cold-paper subset* that drops every test paper present in the Signal-A pool;

Table 9: Overlap between SciRepEval test tasks and the Signal A+B candidate universe (12.4M papers) and the Stage 1 MLM universe (13.4M papers). "Edge" counts positive (q, c) pairs that appear in either direction in the S2AG citation graph restricted to Signal-A candidates; numbers measure unlabelled paper-level exposure only (no downstream labels enter training). *N/A*: external corpus, identifiers not mappable to S2AG.

Task	Kind	#papers	Sig A+B	MLM	#edges	Edge∩
FoS	Classif.	472	7.6%	6.6%	—	—
MeSH Descriptors	Classif.	51,738	13.4%	10.3%	—	—
Biomimicry	Classif.	2,748	20.9%	25.1%	—	—
DRSM	Classif.	954	4.3%	4.1%	—	—
SciDocs MAG*	Classif.	3,709	5.3%	5.6%	—	—
SciDocs MeSH*	Classif.	3,521	18.0%	17.5%	—	—
Citation Count	Regress.	6,012	18.7%	16.8%	—	—
Publication Year	Regress.	6,000	27.5%	18.2%	—	—
h-Index	Regress.	N/A	N/A	N/A	—	—
Peer-Review Score	Regress.	N/A	N/A	N/A	—	—
Tweet Mentions	Regress.	N/A	N/A	N/A	—	—
Same Author	Proximity	15,585	26.2%	33.7%	62,487	1.3%
Highly Infl. Cite.	Proximity	58,453	16.0%	14.1%	8,840	12.2%
SciDocs Cite*	Proximity	26,642	19.1%	11.9%	4,928	7.0%
SciDocs Co-Cite*	Proximity	25,356	19.4%	11.9%	4,948	2.5%
SciDocs Co-Read*	Proximity	20,263	31.0%	20.4%	4,827	3.8%
SciDocs Co-View*	Proximity	28,394	21.5%	14.9%	4,890	2.3%
RELISH	Proximity	162,971	4.2%	3.7%	135,115	0.0%
Search	Ad-hoc	N/A	N/A	N/A	N/A	N/A
NFCorpus	Ad-hoc	N/A	N/A	N/A	N/A	N/A
TREC-CoVID	Ad-hoc	N/A	N/A	N/A	N/A	N/A

*SciDocs tasks use SHA-1 document identifiers; overlap is computed after resolving them to S2 corpus ids via the official meta mapping.

the linear probe is still trained on the full official training set. Proximity tasks are excluded by construction because dropping in-pool test papers would break their query/candidate pairings. Table 10 reports SciEmbed-FULL (single-seed) alongside BGE-large and SPECTER2; all three are run through the same modified evaluator, with SciDocs MAG re-encoded unfiltered as a control whose cold score reproduces the official score to two decimals for every model.

Cold-vs-official deltas are small ($|\Delta| \leq 3.2$) and the signed direction is shared across all three models on every task, regardless of whether they were trained on Signals A+B (SciEmbed), the S2AG citation graph (SPECTER2), or English web text (BGE-large). This is the pattern expected if the delta reflects a property of the cold subset (label-distribution and time-of-publication shift) rather than training-set leakage: a leakage explanation would predict the largest drop for the model trained on Signal-A papers, yet SciEmbed has the smallest mean $|\Delta|$ of the three.

E Recipe and scaling ablations

We run extensive ablations and probes to characterise where SciEmbed's headline number is bounded by supervision vs. recipe.

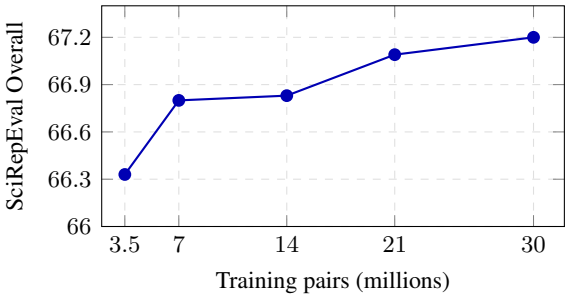


Figure 6: Data-scaling curve (SciEmbed, Signals A+B; all points are seed-42 runs). Gains are monotonic with diminishing returns past 21M pairs (+0.87 total from 3.5M to 30M).

E.1 Data-scaling curve

Figure 6 plots SciRepEval overall against Stage 2 training-pair count from 3.5M to 30M with all other hyperparameters fixed.

E.2 Matryoshka per-seed results

The headline Matryoshka analysis reports SciRepEval overall scores at four truncation dimensions, averaged over three FULL seeds. Table 11 gives the per-seed breakdown so the spread is visible. These overall scores are computed with the `scirepeval` code-path evaluator (as in Table 15), which runs ∼0.3 higher than our main harness; the dim-768

Table 10: Cold-paper subset evaluation on the five highest-overlap classification and regression SciRepEval tasks. Pool is the fraction of test papers in the candidate dataset. F1 macro for classification; F1 (binary) for Biomimicry; Kendall τ_b for Citation Count and Publication Year.

		SciEmbed-FULL			BGE-large			SPECTER2		
Task	Pool	Official	Cold	Δ	Official	Cold	Δ	Official	Cold	Δ
Biomimicry	20.9%	74.6	76.0	+1.5	72.5	75.4	+2.9	74.2	77.4	+3.2
MeSH Descriptors	13.4%	87.3	87.3	0.0	87.9	88.1	+0.1	86.8	86.8	+0.1
SciDocs MeSH	18.0%	89.7	90.0	+0.4	90.1	89.9	−0.2	88.9	89.1	+0.1
Citation Count	18.7%	38.0	36.1	−1.9	36.4	34.4	−2.0	38.4	36.8	−1.7
Publication Year	27.5%	39.6	37.3	−2.3	32.0	30.2	−1.8	35.6	34.0	−1.6
mean $\|\Delta\|$				1.20			1.38			1.33

Table 11: SciRepEval overall ($\times 100$) at each Matryoshka dimension, across three FULL training seeds.

dim	run A	run B	run C	mean ± std
768	67.20	67.10	67.03	67.11 ± 0.08
512	66.88	65.27	65.02	65.72 ± 1.01
256	66.50	64.78	64.94	65.41 ± 0.95
128	66.05	64.20	64.23	64.83 ± 1.06

mean of 67.11 therefore corresponds to the 66.85 that SciEmbed-FULL reaches under the main harness in Table 1.

E.3 Recipe-space ceiling probe

Table 12 groups the 28 single-seed configurations into parameter families, and Table 13 reports the per-task variance underlying the within-SciEmbed vs. across-external-recipe comparison. Probe data-signal construction is detailed in Appendix B.3.

28 configurations. Table 12 groups the 17 contrastive-recipe parameter variants into four parameter families with their 4-category SciRepEval range, and reports the four probe data signals separately as pure-replacement and mixed-with-A+B rows alongside the Phase 3 best-parameter applied to FULL and the SciEmbed-FULL baseline. One additional parameter (`alpha_ep4_lr1e5`, 66.03) is outside the displayed LR/warmup/epoch range and is omitted as an outlier.

Per-task variance: full 13-task table. Table 13 reports within-SciEmbed std (over the 23 single-seed recipe variants in the probe) and across-external-recipe std (over Granite R2, BGE-large, Nomic ModernBERT, SPECTER2 Base, and SPECTER2 + 4 adapters), sorted by external − within gap. Most tasks where SciEmbed is recipe-locked (within std < 1.0) have a meaningfully larger external std, so different supervision sources (not recipes within ours) are what move them; SciDocs MAG and RELISH are the only genuinely benchmark-saturated tasks (low std in both cohorts).

Table 12: Recipe-space ceiling probe (28 single-seed SciEmbed configurations, Granite R2 shown only as external reference). The 17 NoDAPT-CTX parameter variants are grouped into four parameter families and reported as ranges. The best configuration ($1/\tau = 18$ applied to FULL) reaches 67.11, +0.01 above the SciEmbed-FULL baseline (within seed noise). No configuration exceeds the baseline by more than +0.01. Overall scores here use the `scirepeval` code-path evaluator (as in Table 11), ∼0.3 above the main harness of Table 1. n is the number of variants in that row.

Configuration	n	4-cat
Reference		
Granite Embedding R2 (external)	1	67.44
SciEmbed-FULL baseline (seed 123)	1	67.10
Best parameter applied to FULL recipe		
FULL + $1/\tau = 18$	1	67.11
NoDAPT-CTX parameter sweep on 7M-pair A+B		
Contrastive scale $1/\tau \in [12, 30]$	6	66.70–66.82
Loss formulation (MNRL variants)	4	66.48–66.74
LR / warmup / epoch budget	4	66.52–66.68
Pooling / Matryoshka schedule	3	66.36–66.63
DAPT+CTX (Stage 1 + A+B, seed 42)	1	66.17
Additional data-signal sources (mixed with A+B)		
+ intent-filtered cites	1	66.63
+ co-citation pairs	1	66.53
+ TLDR↔abstract pairs	1	65.46
+ OpenAlex related-works	1	65.08
Additional data-signal sources (pure, no A+B)		
Intent-filtered cites only	1	65.38
Co-citation pairs only	1	65.37
TLDR↔abstract pairs only	1	63.60
OpenAlex related-works only	1	63.25

E.4 Task-specific adapter experiments

Per-task adapters trade single-model simplicity for category-specific lift; we evaluate both directions, re-running SPECTER2's released task adapters and training matched LoRA adapters on SciEmbed-

Table 13: Per-task SciRepEval variance, within SciEmbed (23 single-seed recipe variants) vs. across external recipes (Granite R2, BGE-large, Nomic ModernBERT, SPECTER2 Base, SPECTER2 + adapters).

Task	Bucket	SciEmbed	External	Δ
Citation Count	regr.	1.19	4.12	+2.93
NFCorpus	search	1.00	2.99	+1.99
Search	search	0.27	1.60	+1.33
Publication Year	regr.	2.68	4.04	+1.36
Peer Review Score	regr.	1.34	2.55	+1.21
MeSH	cls	0.62	1.79	+1.17
Fields of Study	cls	0.89	2.00	+1.11
SciDocs Cite	prox.	1.30	2.37	+1.07
Same Author Detection	prox.	0.73	1.42	+0.69
TREC-CoVID	search	0.68	1.25	+0.57
Tweet Mentions	regr.	0.46	1.01	+0.55
SciDocs MAG	cls	0.57	0.63	+0.06
RELISH	search	0.78	0.68	−0.10
Bucket mean (cls)	cls	0.78	1.31	+0.53
Bucket mean (regr.)	regr.	1.46	2.60	+1.14
Bucket mean (prox.)	prox.	0.97	1.34	+0.37
Bucket mean (search)	search	0.69	1.63	+0.94

Table 14: SPECTER2 Base vs. SPECTER2 with four task-routed adapters. Δ is the task-routing lift. The matched-configuration number reported by Singh et al. (2023) for the released Adapters checkpoints is 70.9 under their own aggregation (Appendix C.2); per-category directions of effect are consistent.

Category	Base	+ Adapters	Δ
Classification	74.9	**77.0**	+2.1
Regression	27.7	**31.7**	+4.0
Proximity	81.4	**81.8**	+0.4
Search	81.7	**82.7**	+0.9
Overall	66.4	**68.3**	+1.9

FULL.

E.4.1 SPECTER2 with task-routed adapters

The single-model vs. task-routed distinction (§4) is central to our baseline choice, so we reran the official SPECTER2 task-adapter system (`[CLF]`/`[RGN]`/`[PRX]`/`[QRY]` adapters of Singh et al. (2023)) under the exact protocol used for Table 1 rather than relying on a reported number under a different aggregation; Table 14 reports both configurations.

The lift concentrates on the categories the `[RGN]` and `[CLF]` adapters were trained for (Citation Count +8.5 τ_b, Publication Year +7.7, MeSH Descriptors +5.1 F1-macro); proximity is unchanged because the `[PRX]` adapter is the SPECTER2-base weights themselves.

Table 15: SciEmbed-FULL with and without LoRA adapters. Each adapter is evaluated only on its task category. Δ is change from FULL. Bucket averages use the `scirepeval` code path (a different evaluator from our main harness); the seed-42 FULL row gives Overall 67.2 here vs. 66.85 in Table 1 (3-seed mean). Adapter deltas are interpretable within this table.

Category	FULL	+ Adapter	Δ
Classification	75.9	74.5	−1.4
Regression	29.0	**31.9**	+2.9
Proximity	81.0	78.6	−2.4
Search	82.9	81.3	−1.6
Overall	**67.2**	66.6	−0.6

E.4.2 SciEmbed-FULL with LoRA adapters

Following the SPECTER2 multi-adapter paradigm, we trained four LoRA adapters on SciEmbed-FULL, one per task type (CLF/RGN/PRX/QRY). Each adapter is trained for 2 epochs at lr 5×10^{-5}. LoRA weights are merged into the base, yielding four task-specialised checkpoints; Table 15 reports per-bucket scores against FULL.

The regression adapter yields the largest improvement (+2.9), driven by Citation Count (+5.9) and Publication Year (+6.0), the directly supervised tasks. Classification, proximity, and search all degrade: independent per-type LoRA overspecialises at the expense of general representation quality. SPECTER2's task-routed system gains +1.8 overall over its own base under the same setup, plausibly because (i) it uses joint multi-task training with shared routing rather than four independent adapters, and (ii) SciEmbed-FULL's base representations are already stronger (66.85 vs. 66.4), leaving less headroom and making them more sensitive to fine-tuning perturbation. Multi-task LoRA with task routing or AdapterFusion could be a next step.

E.5 Granite R2 initialised with citation-context Stage 2

To test whether citation-context supervision *composes* with a strong general-retrieval backbone rather than only substituting for it, we run our Stage-2 recipe from the Granite R2 weights instead of our Stage-1 MLM checkpoint: identical A+B citation-context data (the 7M-pair CTX mix), identical hyperparameters, with CLS pooling to match Granite's native pooling. Results are the mean±sample-std over three seeds (Table 16).

Fine-tuning Granite R2 on our citation contexts raises its regression from 25.9 to 28.0 ± 0.01

Table 16: Fine-tuning Granite R2 on our open citation-context data (A+B), three seeds. The off-the-shelf Granite R2 and SciEmbed-CTX rows are from Table 1. Citation-context Stage 2 adds Granite's missing regression specialisation (+2.1) while retaining its proximity and most of its search.

Model	Cls.	Regr.	Prox.	Search	Overall
Granite R2 (off the shelf)	76.3	25.9	82.7	84.9	67.4
SciEmbed-CTX (ours)	75.5	28.3	80.9	82.5	66.8
Granite R2 + CTX (ours)	75.6	28.0	82.7	84.2	**67.6**

(close to SciEmbed-CTX's 28.3) while keeping proximity unchanged (82.7) and almost all of its search (84.9 → 84.2). The +2.1 regression gain more than offsets the small classification (−0.7) and search (−0.7) give-back, so the combined model reaches 67.6 ± 0.03 overall (three seeds), above off-the-shelf Granite R2 (67.44) and our own SciEmbed-CTX (66.8). The citation-context signal thus transfers its regression specialisation onto a strong general-retrieval backbone at little cost to the other buckets: the two supervision sources compose rather than trade off, which is the reading the controlled comparison in Section 5.1 argues for, and it shows the specialisation is not tied to our particular Stage-1 initialisation.

F Full MTEB Science evaluation

Table 17 reports SciEmbed performance on nine science-relevant MTEB tasks, evaluated using the official `mteb` package (v2.12.15) under identical conditions for all models shown; no leaderboard estimates.

Within the SciEmbed cohort, FULL improves over CTX on 8/9 tasks (largest gain SciFact +5.0) and CTX over BASE on 7/9 (largest TREC-COVID +5.0, BiorxivClustering +2.2). The 9-task mean inverts the SciRepEval picture: matched-architecture Granite R2 (53.7) leads on five tasks while SciEmbed-FULL reaches 47.0, so general-retrieval supervision dominates MS-MARCO-style scientific retrieval. Granite R2 vs. Nomic ModernBERT differ by 3.6 on the mean here and 1.1 on SciRepEval, so the supervision-source axis dominates over recipe within both benchmarks. SciEmbed-FULL nevertheless leads ArxivClassification (75.5) and MedrxivClustering (37.1) outright, the two tasks closest to citation-context supervision. SciNCL trails SciEmbed-BASE on 7/9 tasks under matched citation-edge supervision but on the SciBERT-512 backbone, consistent with the backbone-upgrade lift in Section 5.1. On the full 41-task MTEB(eng,v2), SciEmbed-BASE/CTX reach 49.9/49.3 vs. Nomic Embed v1.5's 63.4; the gap is expected since SciEmbed's contrastive training uses only citation signals.

G Long-context retrieval

We separate the cost of training at longer Stage-2 context (on SciRepEval) from the benefit (on LongEmbed aggregate and per-length NeedleRetrieval).

G.1 Stage-2 long-context ablation

Table 18 sweeps Stage 2 `max_seq_length` on the CTX recipe: SciRepEval is insensitive to it while LongEmbed (Table 19) is not.

G.2 LongEmbed main table

Table 19 reports aggregate LongEmbed scores for the four SciEmbed-CTX context-length variants alongside the matched-architecture peer Granite R2 and three general-purpose long-context models (up to 3.8× larger). The within-recipe transfer is monotonic (24.7 → 28.9 → 37.0); SciEmbed's citation-context supervision otherwise does not target general long-context retrieval, matching the SciRepEval supervision-source pattern.

G.3 LongEmbed per-task results

Table 20 reports the per-task LongEmbed scores that aggregate to the "LongEmbed avg" column of Table 19. Each value is the mteb `main_score` (nDCG@10 for the four multi-document tasks; Recall@1 for the two single-document tasks) averaged across all eight length splits where the task is multi-length (Needle, Passkey).

Endpoint ordering CTX-512 → CTX-8192 is preserved on all six tasks; strict monotonicity through CTX-2048 holds on three (QMSum, SummScreenFD, Needle). The other three show a CTX-2048 dip that CTX-8192 recovers from: ≤ 0.2 on NarrativeQA and WikimQA (within noise), −5.2 on Passkey before recovering to +6.2 above CTX-512. Granite R2 is co-leader with the much larger BGE-M3 and GTE-large on the four multi-document tasks and leads the two synthetic needle/passkey probes outright; SciEmbed-CTX-8192 sits below the long-context peer group on every task.

Table 17: MTEB science task scores ($\times 100$), all evaluated locally with official `mteb` v2.12.15 under identical conditions; no leaderboard estimates. Bold = best, underline = second-best. GR2 = Granite Embedding R2 (matched-architecture peer); NMB = Nomic ModernBERT; BGE-b/l = BGE-base/large-en-v1.5; Nomic = Nomic Embed v1.5; E5 = E5-base-v2; SPEC2 = SPECTER2 Base. SciDocsRR is the official MTEB v2 main score (`MAP@1k`).

Task	Type	BASE	CTX	FULL	GR2	NMB	SciNCL	BGE-b	BGE-l	Nomic	E5	SPEC2
ArxivClass.	Classif.	73.7	75.1	**75.5**	68.5	72.1	72.1	73.1	74.2	60.7	68.2	73.9
BIOSSES	STS	83.0	78.6	79.0	86.5	85.4	75.3	**86.9**	84.7	86.2	83.6	78.4
SciDocsRR	Rerank	5.9	6.3	6.5	**9.1**	7.2	6.5	8.2	8.5	7.4	7.0	6.5
SciFact	Retrieval	61.6	62.2	67.2	**75.8**	67.6	54.0	73.8	73.5	70.2	71.1	42.8
TRECCOVID	Retrieval	45.4	50.4	50.4	**70.4**	67.8	33.1	67.2	64.4	70.0	52.2	33.7
ArguAna	Retrieval	42.1	41.3	42.7	59.2	50.6	38.4	**63.6**	63.6	53.6	51.6	36.7
NFCorpus	Retrieval	24.9	25.9	26.6	**37.6**	30.0	18.8	36.8	36.9	34.1	35.3	11.3
BiorxivCl.	Cluster	35.4	37.6	37.9	**39.8**	35.7	36.5	39.2	39.7	33.3	35.2	38.4
MedrxivCl.	Cluster	36.2	36.6	**37.1**	36.2	34.6	35.6	35.1	34.5	32.9	32.8	35.6
9-task avg		45.4	46.0	47.0	53.7	50.1	41.1	**53.8**	53.3	49.8	48.6	39.7

Table 18: Long-context Stage 2 ablation. All rows use the SciEmbed-CTX recipe (Signals A+B, 7M pairs, 3 epochs); only `max_seq_length` differs. Longer context yields no SciRepEval improvement (which contains essentially no long inputs). GPU-h is measured single-GPU `train_runtime`; CTX-512 and CTX-2048 ran on H100, CTX-8192 on A100, so wall-times are not directly comparable. The CTX-8192 run was early-stopped at 2.5 epochs to match the $\leq$ 3-epoch budget.

`max_seq_len`	SciRepEval Overall	GPU-h
512	66.8	31
2K	66.5	22
8K	66.5	39

Table 19: Official LongEmbed results via the upstream `mteb` package (v2.12.15). Aggregate is the mean `main_score` over the six benchmark tasks (nDCG@10 for four multi-document tasks, R@1 for the two single-document tasks; multi-length tasks averaged over the eight 256–32K length splits). Best in **bold**.

Model	Params	Ctx	LongEmbed avg
General-purpose long-context			
BGE-M3	568M	8K	59.5
GTE-large-v1.5	434M	8K	52.2
Nomic Embed v1.5	137M	8K	30.5
Matched architecture (149M ModernBERT-8K)			
Granite R2	149M	8K	**65.6**
SciEmbed-CTX	149M	8K	24.7
SciEmbed-CTX-2048	149M	8K	28.9
SciEmbed-CTX-8192	149M	8K	37.0
SciEmbed-NoDAPT-CTX-8192[§]	149M	8K	37.5
SciEmbed-FULL	149M	8K	19.0

[§]Single seed (42). SciEmbed-NoDAPT-CTX-8192 reproduces the CTX-8192 recipe but skips Stage 1 DAPT; the within-recipe match ($\Delta = +0.5$) confirms the SciRepEval NoDAPT finding generalises to long context.

G.4 NeedleRetrieval per length

Table 21 reports Recall@1 by document length on NeedleRetrieval (LongEmbed); the within-recipe transfer summarised in §5.4 is visible at every long length.

H Body-Fact Retrieval

BFR forces the retrieval target into the body: each query is a body sentence with $\leq 20\%$ word-3-gram overlap with its source paper's abstract, so a 512-token model cannot rely on the abstract prefix to recover the paper.

H.1 Construction

BFR is built from S2ORC against the 2026-03-01 snapshot. The candidate pool is 9,749 papers with a non-empty title, an abstract of ≥ 50 characters (`s2ag.abstracts`), an S2ORC body of 8,000–80,000 characters (5th percentile 10,599), and citation count $\in [1, 4]$ to guarantee structural exclusion from Stage-2 training (which required ≥ 5). For each paper we draw one sentence at random from the body middle-third (chars 33%–66%) and reject it if more than 20% of its word-3-grams overlap the abstract's; the first 1,000 surviving sentences are kept as queries. Encoding is *title* . *abstract body*, truncated to each model's native `max_seq_length` (short-context models see only the title/abstract prefix; long-context models see the body); embeddings are L2-normalised, and we report Recall@$\{1, 10, 100\}$ and nDCG@10 against the full 9,749-candidate pool. The 0.20 overlap threshold is heuristic and not tuned on a validation set; BFR should be read as conditional on this construction recipe.

Table 20: Per-task LongEmbed scores (×100). NarrQA = NarrativeQA, SummScr = SummScreenFD. Rows of the same architecture group as in Table 19. Best per task in **bold**, second-best underlined across all nine evaluated models. The right-most "Avg" column matches the "LongEmbed avg" in Table 19 to within rounding.

Model	NarrQA	QMSum	SummScr	WikimQA	Needle	Passkey	Avg
General-purpose long-context							
BGE-M3	**48.8**	36.1	93.9	79.5	39.5	59.0	59.5
GTE-large-v1.5	40.0	37.5	**94.6**	84.1	16.8	40.2	52.2
Nomic Embed v1.5	14.5	15.9	72.3	37.0	17.0	26.2	30.5
Matched architecture (149M ModernBERT-8K)							
Granite R2	47.9	**41.5**	93.6	**85.9**	**42.8**	**82.0**	**65.6**
SciEmbed-CTX	12.6	22.9	45.0	28.7	18.2	21.0	24.7
SciEmbed-CTX-2048	12.5	29.2	60.9	28.6	26.8	15.8	28.9
SciEmbed-CTX-8192	16.7	35.6	76.9	34.8	31.0	27.2	37.0
SciEmbed-NoDAPT-CTX-8192[§]	17.8	37.0	75.7	34.4	32.0	27.8	37.5
SciEmbed-FULL	10.9	22.3	37.1	21.8	12.0	9.8	19.0

Table 21: NeedleRetrieval Recall@1 (×100) by document length in tokens. Granite R2 is the strongest model overall, retaining 32 at 8K and tying SciEmbed-CTX-8192 at 16K. Within the SciEmbed-CTX recipe, longer contrastive context produces strictly monotonic gains at every length from 2K to 16K (8K: $4 \to 6 \to 22$; 16K: $0 \to 6 \to 16$); at 256–1K the ordering is non-strict, plateauing at the short-context ceiling. Either way, the long-context contribution is isolated. Best per length in **bold**.

Model	256	512	1K	2K	4K	8K	16K	32K
BGE-M3	**86**	70	**58**	26	34	18	16	8
Nomic v1.5	58	24	12	10	12	4	10	6
GTE-large	76	28	10	10	4	2	2	2
Granite R2	84	**74**	54	**36**	**36**	**32**	16	10
SciEmbed-CTX	52	34	24	18	12	4	0	2
SciEmbed-CTX-2048	56	46	44	26	26	6	6	4
SciEmbed-CTX-8192	54	46	44	30	28	22	16	8
SciEmbed-NoDAPT-CTX-8192[§]	54	36	42	28	32	26	**24**	**14**

Example query/candidate pair. Table 22 shows a representative pair (query id 6): the body sentence states a quantitative methodological claim (a particle-size threshold above which OIT decreases) that the cited paper's abstract enumerates around but never states, so a 512-token model sees a strictly harder task than an 8K-context model.

H.2 Full table and robustness checks

Consolidated BFR results. Table 23 gathers all BFR results in one place: the within-recipe SciEmbed-CTX rows, the matched-architecture Granite R2, two BM25 lexical controls, and BGE-M3 / Nomic Embed v1.5 for cross-model orientation, across all three pools (full, same-field, and paraphrased-query). The two robustness pools are discussed below.

Same-field pool. To rule out cross-field topical separability as the source of the within-recipe lift, we restrict each query's candidate pool to papers whose first-listed S2AG `s2fieldsofstudy` category matches the source paper (median pool size 398 vs. 9,749 in the full pool; "Same-field" columns of Table 23). The within-recipe ordering is preserved and the CTX-512 → CTX-8192 gap *widens* on every metric: $+4.0 \to +7.3$ R@1, $+5.4 \to +6.5$ R@10, $+4.9 \to +6.9$ nDCG@10. The BGE-M3 vs. SciEmbed-CTX-8192 R@1 gap also shrinks from 5.7 to 0.6, with SciEmbed-CTX-8192 overtaking BGE-M3 on R@10 (82.7 vs. 78.6) and nDCG@10 (0.685 vs. 0.670). We do not present this as a leaderboard claim – same-field BFR carries the same lexical-leakage caveat as the full pool, and the SciEmbed-vs-BGE-M3 swap may be sensitive to the topic-overlap definition – but it does establish that topical separability is not what drives the cross-model gap.

Paraphrased queries. To rule out the dense-side counterpart of the BM25 (t+a+b) leakage path, we rewrite all 1,000 queries with Qwen3-32B (`enable_thinking=False`) under an instruction to preserve facts, numbers, and named entities while changing $\geq 60\%$ of content words (instruction only, not post-hoc filtered); the candi-

Table 22: Example BFR query and its gold candidate (query id 6 in the released file). Both query and gold candidate refer to the same paper, since by construction each query is a body sentence drawn from the paper that is the retrieval target. The trigram-overlap filter rejects queries whose word-3-gram set has more than 20% of its members in the abstract's word-3-gram set, so the answer is in the body but not in the abstract prefix.

Query (body sentence): "In comparisons between samples with different particle sizes, there is a limit above which OIT significantly decreases."
Gold candidate title: How the Dimensions of Plant Filler Particles Affect the Oxidation-Resistant Characteristics of Polyethylene-Based Composite Materials
Abstract excerpt (truncated): "This study analyzed the possibility of using plant-originated waste materials [...] as functional polyethylene fillers. The dependence between the size of the filler particles and their antioxidant potential is discussed. Six fractions were selected: below 50 μm, 50–100 μm, [...] using the extraction method to analyze antioxidant activity (DPPH method) and total phenolic content [...]"
Why the body is needed: the abstract describes *which fractions were tested* and what overall antioxidant assays were used, but the specific quantitative claim that there exists a particle-size threshold above which oxidation-induction time (OIT) drops is reported only in the results section of the body.

Table 23: Body-Fact Retrieval (BFR) diagnostic probe on 9,749 scientific full-text papers (1,000 body-sentence queries), across three pools. *Full*: all 9,749 candidates. *Same-field*: candidates restricted to the source paper's first-listed S2AG `s2fieldsofstudy` category (median 398). *Paraphrased*: queries rewritten by Qwen3-32B to change $\geq 60\%$ of content words, full pool. The probe isolates Stage-2 context length on the SciEmbed-CTX recipe: varying only `max_seq_length` (512 → 8192) is monotonic on every metric in every pool. BM25 (t+a+b) indexes the body, which contains the query verbatim, so its R@1 is the exact-match ceiling, not a comparable dense baseline. **Bold** marks the best dense model per column.

		Full pool			Same-field			Paraphrased		
Model	Ctx	R@1	R@10	nDCG	R@1	R@10	nDCG	R@1	R@10	nDCG
Lexical controls (full-pool reference)										
BM25 (t+a)	512	34.2	47.1	0.407	–	–	–	–	–	–
BM25 (t+a+b)[†]	8K	95.2	98.9	0.971	–	–	–	–	–	–
Cross-model dense (orientation)										
Nomic Embed v1.5	8K	12.7	30.8	0.208	23.3	52.8	0.366	7.8	21.1	0.135
BGE-M3	8K	47.3	64.3	0.553	56.1	78.6	0.670	31.4	45.6	0.385
Granite R2	8K	**50.9**	**72.0**	**0.611**	**61.4**	**85.3**	**0.731**	**36.2**	**51.8**	**0.438**
Within-recipe (SciEmbed-CTX, only `max_seq_len` *differs)*										
SciEmbed-CTX	512	37.6	61.7	0.490	48.2	76.2	0.616	21.7	39.0	0.301
SciEmbed-CTX-2048	2K	39.6	62.2	0.502	50.8	77.3	0.634	22.9	41.7	0.316
SciEmbed-CTX-8192	8K	41.6	67.1	0.539	55.5	82.7	0.685	25.3	45.2	0.347
SciEmbed-NoDAPT-CTX-8192[§]	8K	41.4	66.6	0.535	53.4	82.2	0.675	26.5	46.0	0.357

[†]Contains the query string verbatim; exact-match ceiling, not a dense baseline (paraphrased BM25 not run). [§]Single seed (42).

date pool is unchanged ("Paraphrased" columns of Table 23). Paraphrasing strips $\sim$ 15–17 R@1 from every dense system with a non-trivial baseline (Granite R2 50.9 → 36.2; BGE-M3 47.3 → 31.4; SciEmbed-CTX-8192 41.6 → 25.3), confirming the BM25 ceiling reflected an active leakage path. The within-recipe long-context monotonicity nonetheless survives.

I Efficiency and embedding space

I.1 Inference efficiency

Table 24 and Figure 7 report H100 throughput, latency, and peak VRAM for SciEmbed-FULL and the two strongest baselines, averaged over 50 timed runs (10 warmup) on 300 scientific abstracts via `sentence-transformers`. At `bs=128` in fp16, SciEmbed exceeds BGE-large by 10% at 2.6× less peak VRAM (5,363 vs. 4,875 docs/sec; 445 vs. 1,168 MB); the advantage transfers to a workstation A4500 GPU (1.77× over BGE-large) so it is not datacenter-specific. SPECTER2 leads in absolute throughput but only supports a 512-token input window. The same-backbone Granite R2 is an efficiency wash with SciEmbed (5,375 vs. 5,363 docs/sec at `bs=128`, near-identical single-document latency), so between the two 149M peers the choice is about supervision rather than cost; the throughput advantage is over the 335M-class general models. The higher single-document latency

Model	Prec.	Thr.	Lat.	VRAM
SPECTER2 (110M)	fp32	2700	3.5	1317
BGE-large (335M)	fp32	849	7.3	2303
SciEmbed (149M)	fp32	1926	10.8	858
SPECTER2 (110M)	fp16	8465	3.3	675
BGE-large (335M)	fp16	4875	5.9	1168
SciEmbed (149M)	fp16	**5363**	10.9	**445**
Granite R2 (149M)	fp16	5375	10.8	1200
SciEmbed (149M)	bf16	5277	10.8	445

Table 24: H100 inference benchmark. Throughput in documents/second, latency in ms/doc, peak VRAM in MB at bs=128. Bold marks SciEmbed's primary fp16 deployment configuration; SPECTER2 reaches a higher absolute throughput but is restricted to a 512-token input window and a SciBERT-class encoder. The two 149M ModernBERT peers, SciEmbed and Granite R2, have effectively identical throughput and latency; their peak-VRAM difference reflects checkpoint precision and loading rather than architecture.

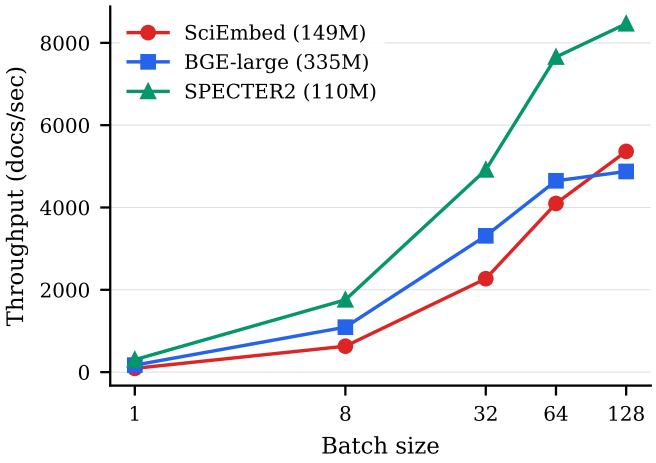


Figure 7: H100 fp16 inference throughput as a function of batch size. SciEmbed narrows the gap to BGE-large from $0.68\times$ at bs=32 to $0.88\times$ at bs=64, then overtakes between bs=64 and bs=128, exceeding BGE-large by 10% at bs=128 while using $2.6\times$ less peak GPU memory.

(10.9 vs. 5.9 ms) reflects ModernBERT's longer kernel-launch path, rarely a bottleneck in batched embedding workloads.

I.2 Embedding-space structure (UMAP)

Figure 8 projects 2,800 Fields of Study papers under UMAP (McInnes et al., 2018) from each of the three matched-architecture embedders. Both general-retrieval models cluster topically adjacent fields together with substantial central overlap; SciEmbed yields cleaner disciplinary separation, with Computer Science forming a distinct cluster and Mathematics, Materials Science, and Chemistry in separate regions. The structure emerges from citation-graph supervision alone (no FoS labels in training), visualising the supervision-source effect that the SciRepEval bucket split in §5.1 reports numerically.

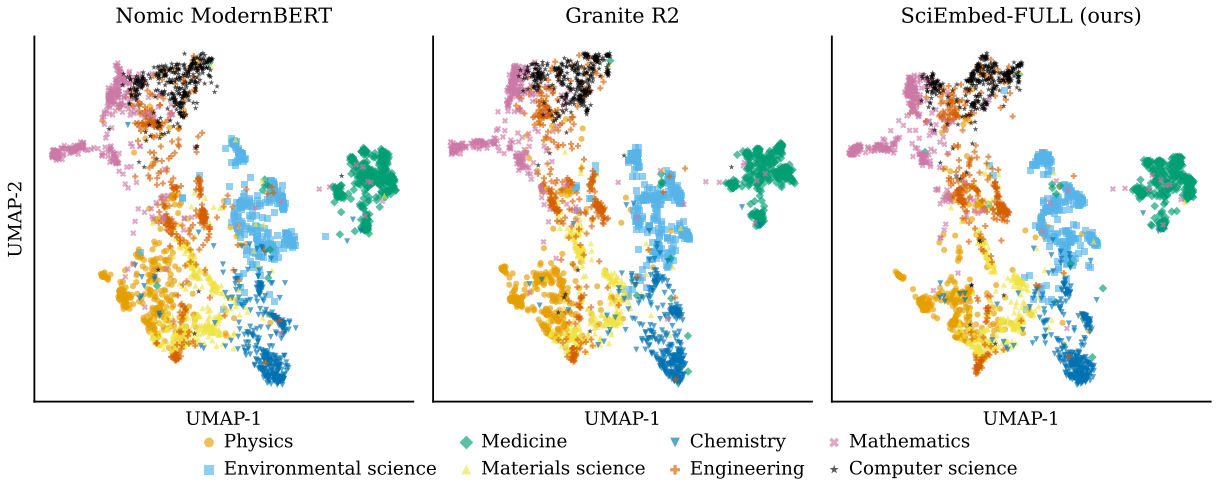


Figure 8: UMAP projection of 2,800 papers from the SciRepEval Fields of Study task (350 per class, top 8 fields) under three matched-architecture (149M ModernBERT-base) embedders with different contrastive supervision: Nomic ModernBERT (general retrieval), IBM Granite R2 (general enterprise retrieval), and SciEmbed-FULL (citation contexts). Each model uses its default pooling (CLS for Granite R2, mean for Nomic ModernBERT and SciEmbed-FULL); UMAP parameters (`n_neighbors=25`, `min_dist=0.15`, cosine metric) are identical across panels. Nomic and Granite produce broadly similar mixed structures; SciEmbed yields visibly cleaner disciplinary separation, consistent with the controlled-comparison finding in Section 5.1.