REIGN: Refurbished Embeddings with Integrated Guidance Networks for Efficient Context-Length Scaling

Devrim Çavuşoğlu[1,2] **Emre Akbaş**[1]

Middle East Technical University[1], OBSS AI[2]

Abstract

Dense retrieval over long documents is expensive. Token-level encoders scale quadratically in sequence length, and most long-context embedding models reach 32K tokens only through architectural workarounds or by stretching billion-parameter LLMs. We propose **REIGN** (Refurbished Embeddings with Integrated Guidance Networks), a contrastively trained bi-encoder that operates on sequences of contextualised chunk embeddings from a frozen Guidance Network (GN) rather than on raw tokens. REIGN targets multi-chunk inputs, primarily for document-to-document retrieval; single-chunk inputs stay with the GN. Decoupling token-level processing from document-level reasoning, and caching the GN embeddings to disk, cuts per-document training cost by roughly four orders of magnitude relative to chunked Transformer fine-tuning. We also release a synthetic long-document retrieval benchmark for contrastive training and evaluation at long context lengths. Across an in-distribution Wikipedia benchmark, the LoCo out-of-distribution suite, and a real-world patent retrieval case study, REIGN matches dense long-context retrievers at smaller parameter budgets in each regime. A paired significance test puts it on par with models 1.6–4.3× larger on the patent task, and it stays within 0.65 nDCG@10 of a 20×-larger model on LoCo. Our code is available at https://github.com/devrimcavusoglu/reign.

1 Introduction

Retrieval-augmented generation, multi-hop question answering, and document-level semantic search all need embeddings for inputs that exceed the context window of a standard Transformer. Efficient-attention variants (Beltagy et al., 2020; Zaheer et al., 2020) and memory-augmented architectures extend the window, but embedding documents that run from tens of thousands to millions of tokens still hits memory and compute walls in practice.

Recent long-context embedding models (e.g., Nomic-Embed, Nussbaum et al., 2025; Jina-Embeddings-v3, Sturua et al., 2025; M2-BERT, Fu et al., 2023) reach 8K–32K tokens via RoPE extensions, sparse attention, or structured sequence mixing. A parallel line instruction-tunes billion-scale LLMs into embedding models (Lee et al., 2025; Choi et al., 2024; Zhang et al., 2025; Meng et al., 2024); these are accurate but computationally heavy, and beyond most groups' training budgets even with parameter-efficient adapters or quantisation. Both lines still take the sub-word token as the base unit of representation.

Data poses a complementary challenge. Existing embedding benchmarks target short passages and lack the semantically aligned long-document pairs and realistic hard negatives required for contrastive learning at scale. We address this with a synthetic long-document retrieval dataset derived from Wikipedia, in which LLM-generated rephrasals plus semantically injected distractors provide positive and graded negative pairs over documents averaging several thousand words (Section 3).

We propose **REIGN** (*Refurbished Embeddings with Integrated Guidance Networks*), a contrastively trained bi-encoder for long-document retrieval. REIGN takes ToBERT's chunk-and-aggregate intuition (Pappagari et al., 2019) and combines it with SimCLR-style contrastive training (Chen et al., 2020). The one shift is the chunk encoder, a frozen pretrained text embedder (a *Guidance Network*, GN; e.g., GTE or E5) whose outputs are cached. REIGN therefore (i) scales to document-level inputs without changing the underlying architecture, and (ii) trains over precomputed chunk embeddings, replacing repeated token-level forward passes with a single small Transformer over a short embedding sequence.

We evaluate REIGN on three retrieval regimes:

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 31659–31677
October 24-29, 2026 ©2026 Association for Computational Linguistics

the in-distribution synthetic Wikipedia benchmark, the out-of-distribution LoCo long-context suite (Saad-Falcon et al., 2024), and a real-world patent prior-art case study (Ayaou et al., 2025). A 55M-parameter REIGN+GTE-small beats every native long-context dense baseline we test on GoodWiki-Long and the LoCo-paper zero-shot baselines on LoCo, and a 357M REIGN+GTE-large is statistically indistinguishable from dense baselines 1.6–4.3× larger on the patent task.

2 Related Work

We situate REIGN with respect to prior work along three axes: (i) representation granularity, (ii) long-context modelling, and (iii) efficiency-oriented training.

2.1 Tokenisation and Representation Granularity

Subword tokenisers such as WordPiece (Wu et al., 2016) and BPE (Sennrich et al., 2016) are the default input layer for Transformer encoders, but at the document scale the resulting token sequences make the attention cost prohibitive.

Modern embedding models such as E5 (Wang et al., 2022), GTE (Li et al., 2023b), and the systems on MTEB (Muennighoff et al., 2023) still treat the subword token as the atomic input. REIGN instead treats the chunk-level embedding from a frozen Guidance Network (GN) as the atomic input, and trains only on top of those embeddings.

2.2 Long-Context Modelling Strategies

Hierarchical encoders split documents into chunks, encode each chunk independently, and aggregate. ToBERT (Pappagari et al., 2019) and CogLTX (Ding et al., 2020) use this recipe for supervised classification over documents of a few hundred to a few thousand words. PARADE (Li et al., 2023a) is the closest precedent for REIGN architecturally. It aggregates passage-level representations from a BERT-style encoder with a small Transformer, but it does so as a query–document cross-encoder for supervised re-ranking. SMITH (Yang et al., 2020), a Siamese hierarchical encoder for document-to-document matching, is the closest in task. Its sentence- and document-level Transformers are trained jointly, need dedicated masked-LM pretraining, and cap inputs at 2,048 tokens. Where these systems train the hierarchy end-to-end, REIGN freezes the chunk encoder and caches its outputs, trains a bi-encoder under a contrastive objective for first-stage retrieval, and targets documents several times longer on average, with tails into the hundreds of thousands of tokens.

A separate line extends the Transformer itself to longer contexts via sparse or structured attention (Longformer and LED (Beltagy et al., 2020), BigBird (Zaheer et al., 2020)) or via sub-quadratic structured-matrix backbones, as in M2-BERT (Fu et al., 2023), which fits an 8K–32K-token encoder into 80M parameters and is the strongest reference on LoCo when trained in-distribution (Saad-Falcon et al., 2024). These approaches change the attention machinery; REIGN keeps a standard Transformer over chunk embeddings and pushes the long-context cost outside the training loop.

The closest settings to ours are therefore SMITH-style Siamese document matching and PARADE-style passage aggregation, but with a frozen, cached chunk encoder, at first-stage retrieval scale, and on documents averaging 5,065 words versus ToBERT's largest 1,788-word corpus (Cieri et al., 2004).

2.3 Efficient and Lightweight Training

Parameter-efficient tuning methods (Adapters (Houlsby et al., 2019), LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), DoRA (Liu et al., 2024)) cut the trainable parameter count when adapting an LLM, but each training step still does a full token-level forward pass over the document. REIGN cuts cost on the other axis. Caching the GN forward passes removes the expensive computation from the training loop entirely, so the trainable encoder only ever sees a short, fixed-length sequence of chunk embeddings.

3 Dataset

3.1 Challenges with Existing Datasets

Existing embedding benchmarks target short-text similarity at the sentence or paragraph level; document-level extensions rarely exceed a few thousand tokens and lack the structural supervision long-context models such as REIGN require. Although large long-form corpora (Wikipedia (Foundation, 2025), Common Crawl (Common Crawl Foundation, 2024), BookCorpus (Zhu et al., 2015), Project Gutenberg (Faysse, 2023)) provide text, they do not contain the semantically aligned long-document pairing required for contrastive learning, and often exhibit noisy formatting or multi-topic

Table 1: Statistics for GOODWIKI-LONG-SYNTHETIC in the BEIR/MTEB IR layout. Per-split rows are query-disjoint qrels, and length stats use `tiktoken cl100k_base`.
score = 2 marks the LLM-rephrased positive, score = 1 a topical distractor.

Pools				
Queries (long-form Wikipedia)				17,854
Corpus (positives + distractors)				53,562
of which positives (score=2, one per query)				17,854
of which distractors (score=1, ≈ 2 per query)				35,708
Qrels splits (query-disjoint)				
Split	Queries	Qrels	score=2	score=1
train	10,000	30,000	10,000	20,000
val	2,000	6,000	2,000	4,000
test	5,854	17,562	5,854	11,708
Document length, tiktoken cl100k_base (median / p95 / max)				
Query side				5,817 / 13,838 / 35,927
Corpus side				1,805 / 2,807 / 32,382

mixing.

3.2 Synthetic Long-Document Dataset

We construct a long-document retrieval dataset by rephrasing long-form Wikipedia articles with GPT-4o-mini (OpenAI, 2024), so each document and its rephrasal form a semantically aligned positive pair that supports contrastive learning over long contexts. We build on GoodWiki (Choi, 2023), a cleaned Wikipedia release in structured markdown, and filter for articles exceeding 16,000 characters (≈4,000 tokens, following OpenAI); we refer to this subset as GOODWIKI-LONG.

To introduce harder negative cases, we augment positives with semantically overlapping distractors inspired by HotpotQA (Yang et al., 2018), yielding a graded relevance structure. During training, fully relevant pairs enter the objective as positives and distractors as partial matches at a reduced target weight (Section 4.2); at evaluation time we use the graded labels to assess ranking robustness.

Open-access contribution. We publicly release GOODWIKI-LONG-SYNTHETIC in the canonical BEIR/MTEB tri-config layout,[1] together with the full training and evaluation code.[2] We refer to this benchmark as GOODWIKI-LONG in the experiments. Pool sizes, qrels splits, and document-length summary statistics are in Table 1; full token-length distributions appear in Figure 1 (`tiktoken cl100k_base`, OpenAI, 2023).

[1] https://huggingface.co/datasets/devrim/goodwiki_long_synthetic_ir

[2] https://github.com/devrimcavusoglu/reign

4 Method

REIGN is trained under a SimCLR-style contrastive scheme (Chen et al., 2020). Each document is transformed into two semantically equivalent views via rephrasing, forming positive pairs (x, x^+); injected distractors act as graded partial matches and other batch samples as negatives. Each view is encoded into chunk-level embeddings by a frozen Guidance Network (GN) and refined by the REIGN encoder under a three-way cosine embedding objective (Section 4.2). Figure 2 diagrams the full pipeline.

4.1 Framework Overview

REIGN treats a pretrained text encoder as a Guidance Network (GN) that maps raw subword tokens to fixed-dimensional chunk embeddings; the REIGN encoder then operates entirely at this coarser embedding level, trading token detail for tractable context length (Figure 3).

A textual input of up to M tokens is segmented by a sliding window of size K (matching the GN's context window, typically $K = 512$) advanced with stride $S \leq K$; consecutive windows overlap by $K - S$ tokens, with $S = K$ giving non-overlapping chunking. Each window C_i is encoded as $E_i = \mathrm{GN}(C_i)$, yielding a sequence $\{E_1, \ldots, E_N\}$ with

$$N = \max\left(1, \left\lceil \frac{M-K}{S} \right\rceil + 1\right), \qquad (1)$$

which reduces to $\lceil M/K \rceil$ when $S = K$. GN outputs are deterministic for a fixed GN/chunk-size/stride/input, so we content-hash and cache them to disk (HDF5); during training only the REIGN encoder runs, over a fixed-length embedding sequence. This compresses the per-document cost from $O(M^2 d)$ at the token level to $O(N^2 d)$ at the chunk level, so lightweight models can encode inputs of 10^5+ tokens; Appendix E formalises what the move to chunk embeddings discards and what training can recover.

The REIGN encoder is a small Transformer applied to the chunk-embedding sequence $\{E_1, \ldots, E_N\}$ as a permutation-equivariant set function with no positional encoding; average pooling across chunks then makes the document representation invariant to chunk order. Prior work and our own ablation support this choice: chunk-level positional encoding gave ToBERT only marginal gains (Pappagari et al., 2019), and in our own ablation learned absolute positions are near parity on all

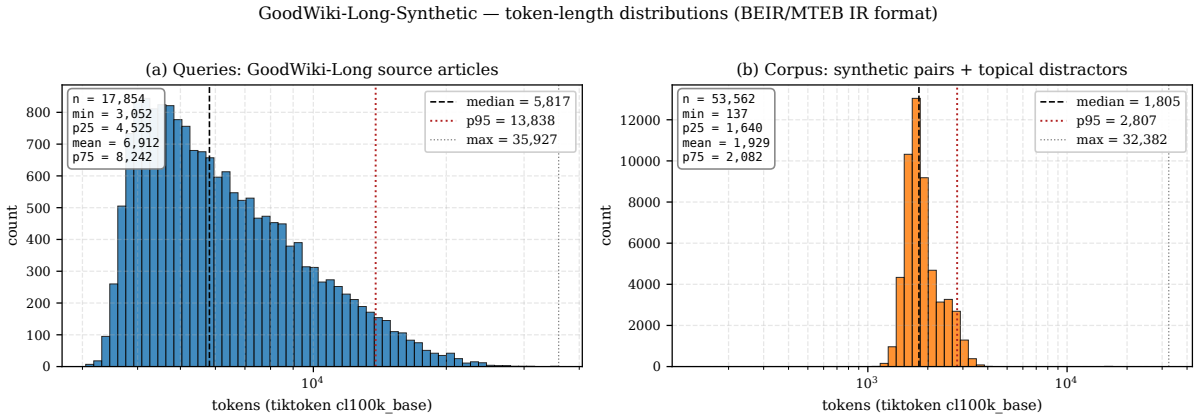

Figure 1: **Token-length distributions of GOODWIKI-LONG-SYNTHETIC** (`tiktoken cl100k_base`). **(a) Queries** (n=17,854) are GoodWiki-Long source articles, long-tailed (median ∼5.8K, p95 ∼13.8K). **(b) Corpus** (n=53,562) holds LLM rephrasals and topical distractors near a 1.8K-token median. The asymmetry puts the long-context burden on the query side, the regime REIGN's cross-chunk encoder targets.

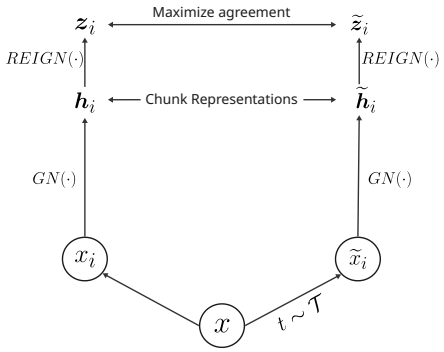

Figure 2: **Training pipeline of REIGN.** Each document x is chunk-encoded by the frozen GN and refined by the REIGN encoder; semantic rephrasing ($t \sim \mathcal{T}$) generates positive pairs, and in-batch negatives enforce separation. Only the positive-pair path is drawn; injected distractors enter the objective as graded partials (Section 4.2).

three benchmarks while sinusoidal encodings are worse (Appendix D).[3] We therefore keep the encoder free of positional parameters. A linear projection adapts the GN output dimension to the encoder hidden width when they differ (e.g. 1024→768 for GTE-LARGE). We sweep depth/width over TINY-L1 (0.56M), SMALL-L2 (3.85M), BASE-L3 (22.45M), LARGE-L4 (52.49M); see Appendix B for the full sweep, and Appendix G for training hyperparameters.

Operating regime. An input that fits a single GN window ($M \leq K$, i.e. N=1) is *short*. Its chunk sequence is one embedding, so the cross-chunk encoder has nothing to aggregate. REIGN's operating regime therefore starts at $N \geq 2$; at K=512 this means inputs beyond one 512-token chunk. Single-chunk inputs go to the bare GN, since on them REIGN trails the GN by 5–7 nDCG@10 (Appendix A) and adds an encoder pass. In deployment REIGN serves the long, multi-chunk side of a corpus alongside a standard short-passage retriever.

4.2 Training Objective

REIGN is optimised with a three-way cosine embedding loss over document pairs (x, y) carrying graded targets $s \in \{1, 0, -1\}$ (positive, partial, negative):

$$\mathcal{L}(x, y, s) = \begin{cases} 1 - c & s = 1, \\ \lambda\,(1 - c) + (1 - \lambda)\,c_+ & s = 0, \\ c_+ & s = -1, \end{cases} \tag{2}$$

where $c = \cos(x, y)$, $c_+ = \max(0, c)$, and the partial weight is $\lambda = 0.5$. Each batch of 18 anchors pairs every anchor with its rephrased positive, its two topical distractors as partials, and 17 in-batch negatives (360 pairs per step; full protocol in Appendix G). The choice of this objective over InfoNCE (Oord et al., 2018) is empirical. In a controlled ablation this objective outperforms every InfoNCE variant we tried (temperature, distractor handling, batch size, warm start, epochs) on all three benchmarks (Appendix H). InfoNCE was also brittle in this setting: at batch size 18 its validation score degrades monotonically after the first epoch, and stable training needs a much larger negative pool (Appendix H). The pairwise

[3] The ablation arms share a controlled protocol (InfoNCE, batch 48, 20 epochs) distinct from the released recipe; see Appendix D.

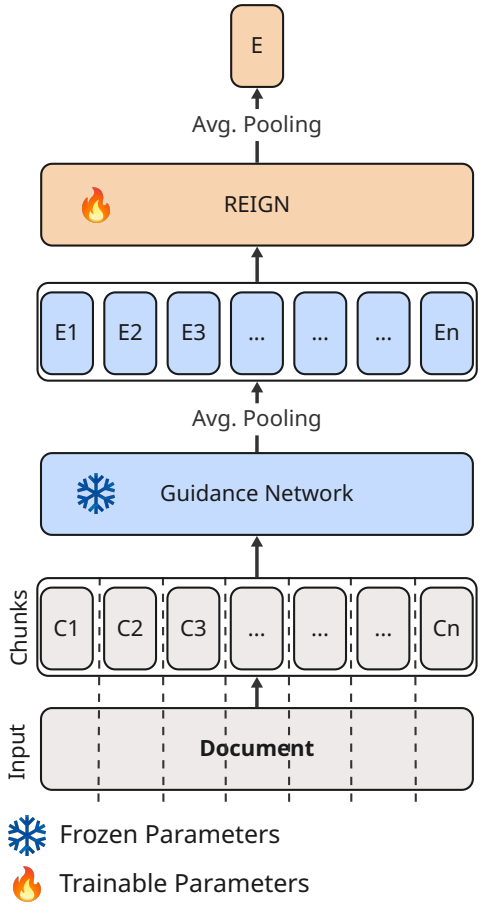

Figure 3: **REIGN architecture.** A long document is chunked; the frozen Guidance Network (GN) encodes each chunk and mean-pools its token embeddings into one contextual chunk embedding $(E_1, \ldots, E_n)$; the trainable REIGN encoder aggregates across chunks and average-pools into a single document representation E.

objective has no such requirement. The DAPFAM fine-tuning experiments (Section 5.3) instead use InfoNCE at temperature 0.07 over the dataset's provided negatives (Appendix I).

4.3 Efficiency

We compare REIGN's cost against representative long-context embedding approaches analytically and empirically, under a fixed 8K-token, 512-token-chunk setting (Appendix F, Table 10). Token-level Transformer models incur substantial overhead, linear in chunk count for chunked inference and quadratic in sequence length for full-context attention. REIGN's cost is instead dominated by a small embedding-level encoder over a short sequence of chunk embeddings. With cached GN embeddings, per-document training cost drops by nearly four orders of magnitude relative to chunked Transformer fine-tuning. Even uncached, REIGN matches standard chunked inference while enabling document-level contrastive training that would otherwise be prohibitive.

End-to-end measurements on a single RTX 4090 agree with this analysis (Table 11 in Appendix F). With cached GN embeddings REIGN answers queries in 0.40–0.52 ms versus 19.8–118.2 ms with the GN run per query (49–229×), and the one-time cache build (8.9–53.3 ms per document) amortises almost immediately. Uncached, REIGN runs at parity with its own chunked GN (ratio 0.86–0.97 over identical GN work). Peak GPU memory spans 0.24–1.73 GB across REIGN configurations, versus 4.8–18.9 GB for the native long-context dense baselines (Jina-v3 at batch 4).

5 Experiments

We evaluate REIGN along three axes: (i) GOODWIKI-LONG (Section 5.1, in-distribution document-to-document retrieval, "doc-to-doc"); (ii) LOCO (Section 5.2, OOD query→passage, 12 subtasks across legal, scientific, and government text); (iii) DAPFAM (Section 5.3, OOD real-world patent doc-to-doc). Primary metrics are nDCG@10/@100 following Thakur et al. (2021); Muennighoff et al. (2023); Saad-Falcon et al. (2024). One shared evaluation implementation scores every system (sparse, dense, and REIGN) over identical qrels; on GOODWIKI-LONG the graded labels enter nDCG with exponential gains $(2^{\text{rel}}-1)$, so every system receives the same partial credit for distractors. A short-context MTEB generalisation study is in Appendix A; REIGN underperforms there by construction (see Limitations). The encoder-capacity ablation that justifies our BASE-L3 paper-default is deferred to Appendix B and Appendix C.

5.1 Retrieval Accuracy on GOODWIKI-LONG

Setup. Table 2 reports retrieval on the GOODWIKI-LONG test split (n_q=5,854, n_{corpus}=53,562) against sparse baselines (BM25, TF-IDF), four native long-context dense baselines (BGE-M3, Chen et al., 2024; Jina-Embeddings-v3, Sturua et al., 2025; Nomic-Embed-v1.5, Nussbaum et al., 2025; Stella-en-1.5B-v5, Zhang et al., 2024), the GN backbones (GTE, Li et al., 2023b; BGE, Xiao et al., 2024) used directly under truncated and chunked mean-pool protocols, and a multi-GN REIGN sweep. Sparse retrievers dominate raw nDCG@10 (BM25 76.82). GOODWIKI-LONG's

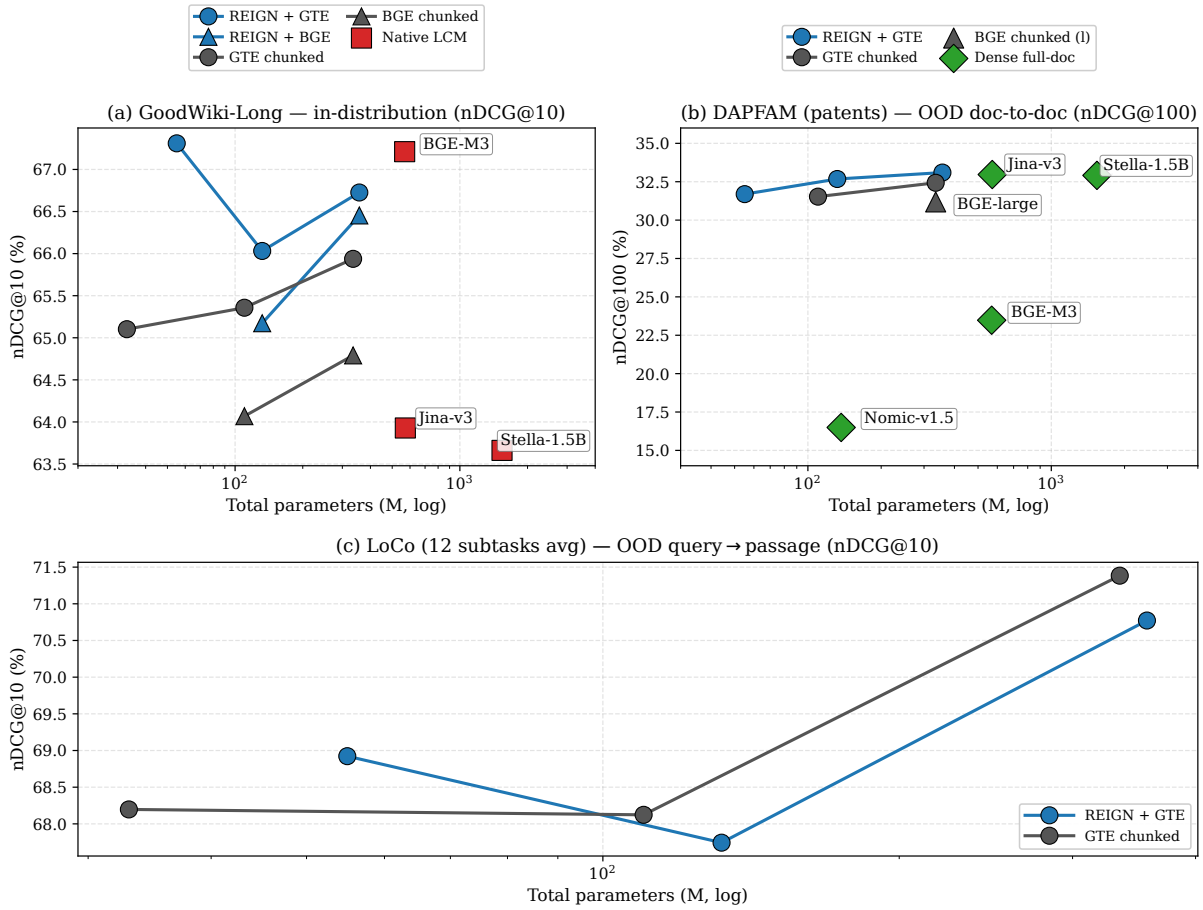

Figure 4: **Parameter efficiency of REIGN across three long-document retrieval regimes.** Total parameters (log-scale) vs. primary nDCG; solid lines connect the three GN scales (small / base / large). **(a) GoodWiki-Long.** REIGN sits above the chunked baseline at every GN scale, and 55M REIGN+GTE-small beats the three plotted native long-context dense baselines (BGE-M3, Jina-v3, Stella-1.5B) at 10–28× smaller scale; Nomic-Embed-v1.5 (11.05 nDCG@10) is omitted for scale. **(b) DAPFAM.** REIGN+GTE matches or exceeds chunked-GTE across scales; the strongest full-document dense baselines (Jina-v3, Stella-1.5B) reach comparable nDCG@100 with 1.6–4.3× more parameters (Table 5). **(c) LoCo.** Bare GTE-chunked is at or above REIGN+GTE except at the small-GN end (see Section 5.2).

positives are rephrased distillations of their queries, and the pairs retain substantial lexical overlap, which favours bag-of-words.

Headline result. Within dense methods, REIGN+GTE-small (55M; Figure 4a) tops nDCG@10, edging BGE-M3 (568M) by 0.10 with 10× fewer parameters and beating Jina-v3 and Stella-1.5B by 3.4–3.7 points, while exceeding the strongest bare-GN chunked baseline (GTE-large, 335M) by 1.4. Chunked mean-pool buys 1.7–2.7 nDCG@10 over truncation across GNs; REIGN's cross-chunk encoder adds a further 0.7–2.2 on top, with the largest lift at the small-GN end.

GN scaling. Within REIGN, GoodWiki-Long performance is nearly saturated at the small-GN end, since scaling the GN from 33M (GTE-SMALL) to 335M (GTE-LARGE) does not improve nDCG@10. REIGN is therefore best paired with the smallest GN that crosses an embedding-quality threshold.

5.2 Out-of-Distribution Long-Context Retrieval: LoCo

Setup. Table 3 reports REIGN zero-shot on LoCo (Saad-Falcon et al., 2024), an OOD suite of 12 long-context retrieval subtasks (legal opinions, scientific QA, government reports, summarisation, code search), each with its own queries, corpus, and qrels. Following the benchmark convention we report per-subtask nDCG@10, and compare systems on the 12-subtask macro-average (Figure 4c). The REIGN encoder is the BASE-L3 checkpoint trained

Table 2: **GoodWiki-Long test.** n_q=5,854, n_{corpus}=53,562, top-k=10. nDCG@10 in percent; Δ is the lift over the truncated baseline of the same GN. REIGN uses the BASE-L3 encoder (22M) at the best-performing stride per GN.
Bold = best per column.
‡Nomic-Embed-v1.5's native training context is 2K tokens, and Nussbaum et al. (2025) mark its 8K RoPE extension as experimental; it collapses on documents whose median length already exceeds 5K.

System	Params	nDCG@10	Δ
Sparse lexical			
BM25	–	**76.82**	–
TF-IDF	–	74.19	–
Native long-context dense (truncate to 8K)			
BGE-M3	568M	67.21	–
Jina-Embeddings-v3	572M	63.93	–
Stella-en-1.5B-v5	1.54B	63.66	–
Nomic-Embed-v1.5 ‡	137M	11.05	–
Bare GN (truncated, 512)			
GTE-small	33M	62.37	–
GTE-base	110M	63.48	–
GTE-large	335M	63.58	–
BGE-base	110M	61.95	–
BGE-large	335M	63.12	–
Bare GN (chunked mean-pool, 512)			
GTE-small	33M	65.10	+2.73
GTE-base	110M	65.36	+1.88
GTE-large	335M	65.94	+2.36
BGE-base	110M	64.07	+2.12
BGE-large	335M	64.79	+1.67
REIGN (this work)			
REIGN + GTE-small	55M	67.31	**+4.94**
REIGN + BGE-base	132M	65.17	+3.22
REIGN + GTE-base	132M	66.03	+2.55
REIGN + BGE-large	357M	66.45	+3.33
REIGN + GTE-large	357M	66.73	+3.15

solely on GOODWIKI-LONG, paired with GTE-SMALL, GTE-BASE, and GTE-LARGE at strides 384 and 512; no LoCo-specific training.

Headline numbers. The bare-GN chunked baseline GTE-SMALL (CHUNKED) already reaches 68.20 macro-average nDCG@10, outperforming BGE-Large (56.5), OpenAI Ada-002 (63.2), and Jina-Embeddings-v2 (67.2). Adding the REIGN encoder on top (55M total) lifts the macro-average to 68.92 (+0.72 over its GN). With a larger GN, REIGN+GTE-LARGE (357M total) reaches 70.77, within 0.65 of the 20×-larger E5-Mistral (Wang et al., 2024) (71.4), and achieves the best zero-shot scores on `passage_retrieval` (64.76) and `qmsum` (64.59).

Where REIGN helps over its GN. REIGN's added value concentrates at the small-GN end. Relative to the same-GN chunked baseline it is +0.72 macro on GTE-SMALL, −0.38 on GTE-BASE, and −0.61 on GTE-LARGE; bare GTE-LARGE (CHUNKED) is the highest LoCo macro-average we observe across any configuration tested (71.38, 0.02 below E5-Mistral at 20× smaller scale). The per-subtask pattern explains the average. REIGN gives large lifts where evidence must be aggregated across chunks (`passage_retrieval` +13.30 at GTE-SMALL, `2wikimqa` +7.48), but a stronger GN's chunked mean-pool already saturates the query→passage subtasks that dominate LoCo (at the bare GN, `gov_report` and `qasper_abstract` are ≥98; `qasper_title` and `summ_screen_fd` are ≥89). REIGN and its GN alike are weak on the lexically driven legal subtasks (`courtlistener_*`, `legal_case_reports`), a term-overlap regime the embedding-level paradigm cannot exploit (see Limitations).

Non-monotone GN scaling. The non-monotone GN-scale pattern (REIGN+GTE-base worse than both small and large) also appears in-distribution (Table 7), so it is not LoCo-specific; we tentatively attribute it to a mismatch between GTE-BASE's chunk-embedding geometry and the fixed 22M REIGN-encoder capacity, and leave a co-varied sweep to future work.

5.3 Real-World Validation: Patent Retrieval on DAPFAM

To test whether REIGN's gains on synthetic long-form Wikipedia carry over to a real specialised domain, we evaluate on DAPFAM (Ayaou et al., 2025), a domain-adaptive patent retrieval benchmark of USPTO patent families. Documents are substantially longer than those in typical retrieval corpora (median ∼7–12K whitespace tokens; tail to several hundred thousand), squarely in REIGN's target regime. We use the FULLTEXT view throughout.

Setup. DAPFAM ships eval-only, so we build a query-disjoint 70/15/15 stratified train/val/test partition (seed 42), holding out n=187 test queries. We retain the authors' IN-/OUT-IPC distinction *within* the held-out set, where TEST_IN (n=180) and TEST_OUT (n=138) are positives-only subsets of TEST, so the cross-domain breakdown is reported on identical queries across systems. Retrieval is against the full 45,336-document FullText corpus with self-matches removed; the authoritative metric is nDCG@100. Fine-tuning uses the dataset's random `score`=0 negatives, with the objective's partial branch disabled (further details in Appendix I).

Table 3: **LoCoV1 per-subtask nDCG@10.** Baselines reproduced from Saad-Falcon et al. (2024), Table 13. "Ctx" is the encoder's native context window.
$\infty^{\dagger}$ marks REIGN (chunked at 512, effective context bounded only by GPU memory).
Above the midrule: zero-shot retrievers, **bold** = best per column. M2-BERT below it was trained on data including the LoCo training splits (Saad-Falcon et al., 2024), so it is shown for reference rather than as a like-for-like comparison; its large legal-subtask margins come from in-distribution supervision.
Abbreviations: 2wMQA, CL_H/T, GovR, LCR, MFQA, PR, QasA/T, QMS, SO, SSFD = 2wikimqa, courtlistener_HTML/Plain, gov_report, legal_case_reports, multifieldqa, passage_retrieval, qasper_abstract/title, qmsum, stackoverflow, summ_screen_fd.

System	Params	Ctx	2wMQA	CL_H	CL_T	GovR	LCR	MFQA	PR	QasA	QasT	QMS	SO	SSFD	Avg
BGE-Large	335M	512	69.4	10.0	10.1	92.6	18.9	89.3	20.1	94.5	87.1	46.0	74.1	65.8	56.5
OpenAI Ada-002	–	8K	78.9	16.3	16.8	97.1	28.2	90.1	31.2	98.9	93.8	48.8	72.3	86.2	63.2
Jina-Embeddings-v2	137M	8K	81.6	27.0	26.1	98.6	30.7	86.4	60.7	99.4	95.1	40.5	69.0	93.3	67.2
E5-Mistral	7.11B	4K	**88.3**	**33.9**	**34.6**	98.3	**49.5**	**93.5**	35.3	**99.8**	**98.4**	46.8	**82.7**	**95.9**	**71.4**
GTE-small (chunked)	33M	512	78.67	28.99	26.68	98.67	33.83	88.41	49.88	99.67	93.76	64.48	61.70	93.61	68.20
GTE-base (chunked)	110M	512	82.32	27.91	26.53	**98.76**	31.81	91.14	50.92	99.32	94.29	63.36	62.06	89.07	68.12
GTE-large (chunked)	335M	512	85.78	31.16	29.28	98.45	39.75	92.06	62.76	99.70	96.73	63.12	63.56	94.25	71.38
REIGN + GTE-small @s384	55M	$\infty^{\dagger}$	79.65	27.23	26.41	98.55	34.26	88.19	63.18	98.49	90.19	63.08	63.46	93.75	68.87
REIGN + GTE-small @s512	55M	$\infty^{\dagger}$	86.15	26.15	24.89	98.67	32.69	89.45	60.53	98.69	90.11	63.37	62.69	93.68	68.92
REIGN + GTE-base @s384	132M	$\infty^{\dagger}$	81.68	24.73	24.41	98.41	29.12	87.56	60.19	98.72	90.72	60.48	65.45	91.46	67.74
REIGN + GTE-base @s512	132M	$\infty^{\dagger}$	80.96	24.86	23.67	98.53	30.07	88.93	56.39	98.97	91.22	60.98	65.25	92.79	67.72
REIGN + GTE-large @s384	357M	$\infty^{\dagger}$	83.90	27.69	26.70	98.62	40.11	88.36	**64.76**	99.47	94.21	**64.59**	67.37	93.49	70.77
REIGN + GTE-large @s512	357M	$\infty^{\dagger}$	84.94	27.56	26.14	98.74	39.98	89.96	61.18	99.56	94.93	63.00	66.81	93.09	70.49
Trained on LoCo splits; not a zero-shot comparison.															
M2-BERT-8K	80M	8K	87.7	94.9	95.4	96.5	58.6	97.1	82.0	98.9	93.1	69.6	84.8	84.6	86.9
M2-BERT-32K	80M	32K	92.1	96.8	97.0	98.7	67.1	98.5	89.3	98.3	97.9	70.4	88.9	98.0	91.1

Headline result. Table 4 compares REIGN against sparse (BM25, TF-IDF), native long-context dense (BGE-M3, Jina-Embeddings-v3, Nomic-Embed-v1.5, Stella-en-1.5B-v5), and chunked dense (BGE-large, GTE-small/base/large) baselines. REIGN trained *only* on GOODWIKI-LONG, with no patent-specific adaptation, has the best mean, 33.10 nDCG@100 (REIGN+GTE-LARGE, 357M total; Figure 4b), and is statistically indistinguishable from the strongest single-stage dense baselines Jina-v3 (572M) and Stella-1.5B (1.54B) with 1.6–4.3× fewer parameters (Table 5); the lift over GTE-LARGE chunked alone is $+0.67$. nDCG@100 grows monotonically with GN size ($31.69 \rightarrow 32.72 \rightarrow 33.10$, best stride per GN).

Statistical significance. We test the TEST-split margins with a paired bootstrap (B=10,000, seed 42) for confidence intervals and a paired randomisation test (B=10,000) for p-values, Holm-adjusted across the five comparisons (Table 5). All three dense margins are non-significant. REIGN is at statistical parity with Jina-v3 and Stella-1.5B while using 1.6–4.3× fewer parameters, and against both it wins on fewer queries than it loses (81/95 and 80/95); its higher mean comes from larger gains on the queries it wins. Against its own chunked GN the margin of $+0.67$ is a boundary case, since the bootstrap CI marginally excludes zero ($[+0.01, +1.34]$) while the randomisation test does not reach significance (p=.051; Holm-adjusted .154). REIGN is also the *larger* system in that comparison (357M vs. 335M), so the parameter-efficiency framing does not apply there. REIGN beats both sparse baselines significantly; sparse retrieval leads on the synthetic GOODWIKI-LONG (Section 5.1) but not on these human-written patents.

Cross-domain decomposition. The TEST_IN/TEST_OUT columns split the held-out set into IN-IPC and cross-IPC positives. Absolute nDCG is dominated by IN-IPC (∼38%); the cross-IPC partition plateaus in the 4.6–5.6 range across the full GTE scale axis (chunked small/base/large) and every REIGN variant. This plateau is an intrinsic property of the OUT-split (random non-cited families provide too weak a training contrast against a topically dense 45K corpus, a limitation noted by the dataset authors), not a model-side deficit. On IN-IPC, REIGN's lift over the matched bare-GN chunked baseline narrows with GN scale ($+2.05$ at GTE-SMALL, $+0.97$ at GTE-BASE, $+0.65$ at GTE-LARGE), mirroring the in-distribution GOODWIKI-LONG pattern.

Fine-tuning does not exceed zero-shot. To rule out under-tuning, we swept fine-tuning configurations over warm- vs. cold-start, backbone scale, stride, and a learning-rate × weight-decay grid (Ap-

Table 4: **Patent retrieval on DAPFAM.** nDCG@100 on TEST (n=187), IN-IPC subset TEST_IN (n=180), and cross-IPC subset TEST_OUT (n=138); 45,336-doc FullText corpus. Δ is the lift on TEST over the same-GN bare chunked baseline (in/out columns not evaluated for sparse and full-document baselines). REIGN rows use the GOODWIKI-LONG-trained zero-shot BASE-L3 checkpoint at the best-performing stride per GN.
Bold = best mean per column; on TEST the margins between REIGN and the strongest dense baselines are not statistically significant (Table 5).
‡Nomic-Embed-v1.5's native training context is 2K tokens, and Nussbaum et al. (2025) mark its 8K RoPE extension as experimental; it collapses on documents whose median length already exceeds the native window.

System	Params	test	test_in	test_out	Δ
Sparse lexical					
BM25	–	25.62	–	–	–
TF-IDF	–	30.06	–	–	–
Native long-context dense (truncate to 8K)					
BGE-M3	568M	23.48	–	–	–
Jina-Embeddings-v3	572M	32.97	–	–	–
Nomic-Embed-v1.5 ‡	137M	16.49	–	–	–
Stella-en-1.5B-v5	1.54B	32.91	–	–	–
Bare GN (chunked mean-pool, 512)					
GTE-small	33M	29.95	34.92	4.63	–
GTE-base	110M	31.53	36.68	5.14	–
GTE-large	335M	32.43	37.50	5.22	–
BGE-large	335M	31.19	–	–	–
REIGN (this work)					
REIGN + GTE-small	55M	31.69	36.97	4.95	**+1.74**
REIGN + GTE-base	132M	32.72	37.65	**5.55**	+1.19
REIGN + GTE-large	357M	**33.10**	**38.15**	5.02	+0.67

Table 5: **Paired significance on DAPFAM test.** REIGN+GTE-LARGE @s512 (33.10) vs. each baseline on per-query nDCG@100 (n=187). Paired bootstrap 95% CI (B=10,000) and paired randomisation p (B=10,000), Holm-adjusted across the five comparisons (p_{H}). Deltas are computed on unrounded per-query means from the significance re-runs, whose provenance is given in Appendix I.
W/L/T = per-query wins/losses/ties.

Baseline	Δ	95% CI	p	p_{H}	W/L/T
Jina-v3	+0.12	[−1.13, +1.36]	.857	1.000	81/95/11
Stella-1.5B	+0.09	[−1.71, +1.98]	.921	1.000	80/95/12
GTE-large chunked	+0.67	[+0.01, +1.34]	.051	.154	103/73/11
BM25	+7.48	[+6.03, +9.03]	<.001	<.001	136/43/8
TF-IDF	+3.03	[+1.50, +4.60]	<.001	<.001	103/74/10

pendix I). The best run, a regularised warm-start at lr 10^{-6} and wd 10^{-2}, reaches 32.73 nDCG@100, statistical parity with the zero-shot checkpoint it started from (32.68 → 32.73). Naive fine-tuning at lr 10^{-5} degrades by 0.4–1.5 points via catastrophic forgetting (warm) or overfit (cold); cold-start loses an additional 0.43 to warm-start at matched hyperparameters (32.30 vs. 32.73), confirming that the GOODWIKI-LONG-trained encoder carries useful inductive bias outside its training domain.

6 Conclusion

We presented **REIGN**, a framework for scalable document-level representation learning that operates directly on precomputed chunk embeddings from a frozen Guidance Network and trains a lightweight cross-chunk encoder under a contrastive objective. The design (i) scales to inputs of 10^5+ tokens without architectural modifications, (ii) reduces per-document training and inference cost by orders of magnitude through cached GN embeddings (Section 4.3), and (iii) transfers zero-shot from synthetic long-form Wikipedia to a specialised long-document domain (IN-IPC patent retrieval). Training runs on GOODWIKI-LONG-SYNTHETIC, a synthetic long-document retrieval benchmark we release, and Appendix E separates the information the frozen GN discards from what training can recover.

The smallest REIGN configuration (55M, GTE-small) outperforms the LoCo-paper zero-shot baselines (BGE-Large, Jina-v2, Ada-002), the first two at 2.5–6× smaller scale, and REIGN+GTE-large reaches 70.77 macro nDCG@10 within 0.65 points of E5-Mistral with 20× fewer parameters. On DAPFAM REIGN has the best mean overall (33.10 nDCG@100) and on the IN-IPC partition, and is statistically indistinguishable from the strongest single-stage dense baselines Jina-v3 (572M) and Stella-1.5B (1.54B) with 1.6–4.3× fewer total parameters (Table 5); the cross-IPC plateau (∼5 nDCG@100 across every system) is a dataset property.

The LoCo picture is task-conditional. REIGN delivers large lifts on multi-chunk-aggregation subtasks, but the bare GTE-large chunked mean-pool is the Pareto-best LoCo configuration overall. The in-distribution / OOD encoder-capacity ablation (Appendix B, Appendix C) establishes BASE-L3 as the principled default and surfaces a methodological caveat. In-distribution sweet-spot claims must be verified out-of-distribution.

Natural future directions include adaptive or overlapping chunking, joint GN fine-tuning, and hybrid composition with a sparse lexical channel.

Limitations

Task-conditional value over the GN. REIGN's added value over its frozen Guidance Network

scales with how much cross-chunk reasoning the retrieval task actually requires. On doc-to-doc retrieval the lift is consistent. REIGN+GTE-small (55M) tops every dense baseline on GOODWIKI-LONG (Section 5.1), and REIGN+GTE-large (357M) has the best DAPFAM mean against 1.6–4.3×-larger dense baselines (Section 5.3); the margins are small (0.10 nDCG@10 over BGE-M3 in-distribution) and on DAPFAM not statistically significant (Table 5), so the advantage is parameter efficiency rather than raw accuracy. On query→passage long-context retrieval (LoCo, Section 5.2) the chunked mean-pool of a stronger GN already saturates the per-subtask ceiling, and REIGN's encoder adds little on GTE-BASE (-0.38 macro) and slightly hurts on GTE-LARGE (-0.61 macro). On short-passage retrieval where the chunked architecture cannot be exercised at all (MTEB ARGUANA, FIQA-2018; Appendix A), performance drops by 5–7 nDCG@10. Across the three regimes, REIGN's encoder is principally a doc-to-doc cross-chunk-reasoning module, and deployment should target tasks whose retrieval signal requires aggregating evidence across multiple chunks of a document.

Short-context retrieval. REIGN is designed for long-document retrieval and underperforms when inputs are shorter than the chunk size. On the two MTEB short-text benchmarks we evaluated (ARGUANA, FIQA-2018; Appendix A), REIGN trails the truncated GTE-small baseline by 5–7 nDCG@10 points. With a single chunk embedding there is nothing to aggregate across, so the cross-chunk signal REIGN is built to exploit disappears.

Lexically driven long-form text. A concrete in-distribution case appears on the legal subtasks of LoCo (Section 5.2), where `courtlistener_HTML` (26.15 nDCG@10), `courtlistener_Plain` (24.89), and `legal_case_reports` (32.69) trail REIGN's QA and summarisation subtasks by as much as 74 points. These corpora are dominated by highly repetitive legal phrasing, so the retrieval signal is concentrated in term-overlap rather than document-level semantics; a cross-chunk encoder over pooled GN embeddings discards exactly the surface-level lexical contrast that such corpora rely on. We view this as a deliberate trade-off of the embedding-level paradigm; hybrid configurations composing REIGN with a sparse lexical channel are a natural direction for future work.

Coupling to the Guidance Network. REIGN's representation quality is upper-bounded by its frozen GN, so any biases, semantic artefacts, or compression errors in the GN embeddings are inherited unchanged. Switching to a new GN requires full retraining of the REIGN encoder; the cached chunk-embedding artefact is GN-specific. The encoder-capacity ablation (Appendix B, Appendix C) further shows that an in-distribution "sweet-spot" configuration can be misleading, because the OOD-justified default differs from the in-distribution optimum. Practitioners adapting REIGN to a new domain should plan an OOD verification step.

Chunk-size × memory trade-off. Smaller chunks yield finer-grained signals but raise GPU memory usage with chunk count. As a concrete reference, a REIGN-SMALL-L2 model paired with GTE-large at effective batch size 312 consumes approximately 1.8, 3.6, 4.5, and 8.4 GB for chunk sizes 512, 256, 128, and 64 respectively.

Token-level fidelity. Because REIGN consumes abstracted chunk embeddings, it is not suited to tasks that demand token-level fidelity (syntactic parsing, numerical reasoning, fine-grained entity matching), where the lossy compression at the GN boundary discards information the downstream task cannot reconstruct.

Ethics Statement

GOODWIKI-LONG-SYNTHETIC derives from GoodWiki (Choi, 2023), a cleaned release of English Wikipedia distributed under CC BY-SA; our dataset release preserves that licensing and attribution. The rephrased corpus documents are machine-generated with GPT-4o-mini (OpenAI, 2024) and are marked as synthetic in the dataset card. DAPFAM (Ayaou et al., 2025) is built from public USPTO patent records. No human subjects were involved, and neither corpus contains private or personally identifying information beyond what is already public in Wikipedia articles and patent filings.

Acknowledgements

This work was supported by OBSS under project code ARGEM-024.

References

Iliass Ayaou, Denis Cavallucci, and Hicham Chibane. 2025. DAPFAM: A domain-aware family-level dataset to benchmark cross-domain patent retrieval. *Preprint*, arXiv:2506.22141.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

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, Bangkok, Thailand. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 1597–1607. JMLR.org.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy yong Sohn. 2024. Linq-embed-mistral technical report. *Preprint*, arXiv:2412.03223.

Euirim Choi. 2023. Goodwiki dataset. https://www.github.com/euirim/goodwiki.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Common Crawl Foundation. 2024. Common crawl corpus: Cc-main-2024-10. https://data.commoncrawl.org/crawl-data/CC-MAIN-2024-10/index.html. Accessed: 2025-05-02.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 10088–10115, Red Hook, NY, USA. Curran Associates Inc.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. CogLTX: Applying BERT to long texts. In *Advances in Neural Information Processing Systems*, volume 33, pages 12792–12804. Curran Associates, Inc.

Manuel Faysse. 2023. Project gutenberg dataset. https://huggingface.co/datasets/manu/project_gutenberg. Accessed: 2025-05-02.

Wikimedia Foundation. 2025. Wikimedia downloads. https://dumps.wikimedia.org.

Daniel Y. Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W. Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. 2023. Monarch Mixer: A simple sub-quadratic GEMM-based architecture. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 77546–77603, Red Hook, NY, USA. Curran Associates Inc.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. Gemini embedding: Generalizable embeddings from Gemini. *Preprint*, arXiv:2503.07891.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2023a. PARADE: Passage representation aggregation for document reranking. *ACM Trans. Inf. Syst.*, 42(2).

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, pages 32100–32121. JMLR.org.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-Mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

OpenAI. 2023. tiktoken: A fast BPE tokenizer for use with OpenAI's models. `https://github.com/openai/tiktoken`. Version 0.9.0.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Understanding and counting tokens. OpenAI Help Center; accessed 2026-08-28.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Jon Saad-Falcon, Daniel Y Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. Benchmarking and building long-context retrieval models with LoCo and M2-BERT. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42918–42946. PMLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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, Berlin, Heidelberg. Springer-Verlag.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

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. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

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*, SIGIR '24, pages 641–649, New York, NY, USA. Association for Computing Machinery.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 1725–1734, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 17283–17297, Red Hook, NY, USA. Curran Associates Inc.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and Stella: Distillation of SOTA embedding models. *Preprint*, arXiv:2412.19048.

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. *Preprint*, arXiv:2506.05176.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 19–27, USA. IEEE Computer Society.

A MTEB Short-Context Generalisation

Table 6 reports retrieval performance on two MTEB benchmarks (ARGUANA and FIQA-2018) consisting of short queries paired with comparatively short documents. REIGN underperforms the truncated GTE-small baseline by 5.1 nDCG@10 on ARGUANA and 6.8 nDCG@10 on FIQA-2018. The drop is not a GN artefact, since the truncated baseline sees the very same single chunk embedding. It comes from the REIGN encoder, which is trained on multi-chunk sequences: applied to a lone chunk embedding it has nothing to aggregate, and the cross-chunk mechanism it relies on cannot contribute. REIGN targets long-context retrieval, and short-passage tasks lie outside the intended operating regime.

B Encoder-Capacity Ablation: In-Distribution Sweep

Table 7 reports the full 4×3 sweep of REIGN encoder configurations against GTE Guidance Networks on the GOODWIKI-LONG test set, evaluated at top-$k = 10$. The Δ column makes the inverted-U pattern immediately readable. Under-capacity TINY-L1 sits below the bare-GN baseline on all three GNs; over-parameterised LARGE-L4 drops below baseline on GTE-BASE and GTE-LARGE; SMALL-L2 reaches the per-GN maximum on GTE-SMALL and GTE-LARGE, with BASE-L3 essentially tied on GTE-BASE (visualised in Figure 5, left). Read in isolation, this would make SMALL-L2 the paper default. The out-of-distribution head-to-head in Appendix C reverses that conclusion, since BASE-L3 wins 11 of 12 cells (the 12th is a -0.20 near-tie), with average margins of $+0.71$ nDCG@100 on DAPFAM and $+0.59$ nDCG@10 on LoCo. The in-distribution SMALL-L2 optimum is therefore an overfitting artefact; BASE-L3 generalises better under cross-domain transfer, the regime the paper targets, so an in-distribution capacity ablation alone is insufficient. We adopt BASE-L3 as the paper-default REIGN encoder throughout the headline tables.

C Encoder-Capacity Ablation: OOD Head-to-Head Detail

Table 8 reports the full per-cell out-of-distribution head-to-head between SMALL-L2 and BASE-L3 on DAPFAM and LoCo at both strides; the bar-chart panel in Figure 5 (Appendix B) visualises the DAPFAM @s512 slice. BASE-L3 wins 11 of 12 cells (the 12th is a -0.20 near-tie, LoCo at GTE-LARGE stride 384); average margins are $+0.71$ nDCG@100 on DAPFAM and $+0.59$ nDCG@10 on LoCo.

D Positional-Encoding Ablation

We ablate the chunk-position signal referenced in Section 4.1, comparing no positional encoding (our design) against learned absolute chunk positions and against fixed sinusoidal encodings, both injected after the GN projection. We train all three arms from scratch under one protocol, using BASE-L3 on GTE-SMALL, chunk/stride 512, InfoNCE (τ=0.07, partial weight 0.5), batch 48 with 47 in-batch negatives, 20 epochs, lr 10^{-5}, seed 42, and best-validation selection. The arms sit below the released operating point (Table 2) because the released checkpoints use the cosine recipe of Section 4.2.

Table 9 supports the no-PE design. Learned absolute positions are near parity on all three benchmarks, and sinusoidal encodings are worse everywhere; positional encoding adds parameters without improving any benchmark. Unit tests give an independent check. Without a positional signal the pooled document representation is invariant to chunk order (maximum deviation 3×10^{-8}), while the other two arms are order-sensitive.

E Theoretical Analysis: Information Loss and Recovery

We formalise the information lost in the move from sub-word tokens to GN embeddings, and separate that fixed loss from the trainable remainder.

Table 6: **Selected MTEB IR results.** Baselines are evaluated with truncation (512 tokens). REIGN uses the paper-default BASE-L3 encoder (22M) with GTE-SMALL as the Guidance Network. Δ is computed with respect to the truncated GTE-small baseline. ARGUANA provides only a test split.
†Trained only on GoodWiki-Long; no MTEB-specific fine-tuning.

Model	MAP@1	MAP@10	R@10	P@10	nDCG@10	Δ nDCG@10
ArguAna						
GTE-small (baseline)	**30.73**	**46.60**	**83.57**	**8.36**	**55.42**	–
REIGN + GTE-small†	26.53	41.52	78.66	7.87	50.32	−5.10
FiQA-2018						
GTE-small (baseline)	**19.34**	**31.72**	**45.63**	**11.16**	**39.31**	–
REIGN + GTE-small†	15.23	25.66	38.14	9.35	32.55	−6.76

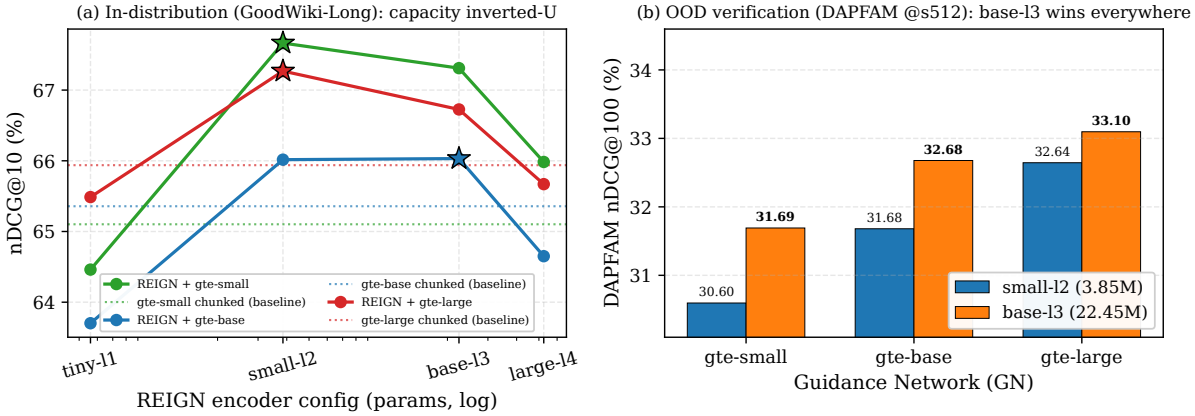

Figure 5: **Encoder-capacity sweet-spot reverses between in-distribution and OOD.** The left panel plots REIGN nDCG@10 on GOODWIKI-LONG vs. encoder size, per GN; stars mark per-GN maxima, dotted lines are bare-GN baselines. The right panel compares DAPFAM @s512 SMALL-L2 against BASE-L3, which wins at every GN scale. Full per-cell numbers in Appendix C.

Scope. Theorem 1 assumes nothing about θ or about the objective that produced it, so it covers every REIGN variant, the released checkpoints included, and Proposition 2 states when the trainable part of the gap closes. Only the recovery bound of Proposition 3 depends on the objective. It covers the InfoNCE-trained arms of Appendices D and H, whose in-batch negatives are draws from the positive marginal as the proposition assumes; the DAPFAM fine-tuning runs (Appendix I) also use InfoNCE, but their provided corpus negatives and false-negative masking sit outside that sampling assumption. The three-way cosine objective of the released checkpoints (Section 4.2) admits no such bound; Remark 4 states what it does give, with the empirical case in Appendix H.

Theorem 1 (Embedding Refurbishment)**.** *Let X be a random variable over natural-language texts, $T = f(X)$ the (deterministic[4]) sub-word token sequence with $H(T) < \infty$ (which holds whenever document length is bounded), $E = g(T)$ the frozen GN embedding sequence, and $R_\theta = h_\theta(E)$ the fixed-size embedding produced by a REIGN encoder h_θ with trainable parameters θ. Because f and g are deterministic, $X \to T \to E \to R_\theta$ is a Markov chain and T, E, R_θ take countably many values. For every θ and every measurable h_θ:*

(i) (GN coarsening.) *$I(X;E) \le I(X;T)$, and the gap is*

$$I(X;T) - I(X;E) \;=\; H(T \mid E),$$

which vanishes iff g is injective on the support of T.

(ii) (End-to-end loss decomposition.) *The total loss in mapping X to R_θ decomposes as*

$$I(X;T) - I(X;R_\theta) \;=\; \underbrace{I(X;T) - I(X;E)}_{\text{GN coarsening (fixed)}} + \underbrace{I(X;E) - I(X;R_\theta)}_{\text{closed by training}},$$

[4]For a fixed tokeniser and vocabulary.

Table 7: **REIGN config × GN ablation on GOODWIKI-LONG test (in-distribution).** nDCG@10 (×100); Δ measured against the per-GN bare chunked-mean-pool baseline. Head dimension is held at d/H=64 across configs. $L/d/H$/FFN = the REIGN encoder's number of Transformer layers, hidden width, attention heads, and feed-forward dimension.
Bold = best REIGN config per GN.

REIGN config	Params	L	d	H	FFN	GTE-small (Δ)	GTE-base (Δ)	GTE-large (Δ)
GN-only baseline	–	–	–	–	–	65.10	65.36	65.94
TINY-L1	0.56M	1	192	3	768	64.46 (−0.64)	63.70 (−1.66)	65.49 (−0.45)
SMALL-L2	3.85M	2	384	6	1536	**67.67** (+2.57)	66.02 (+0.66)	**67.27** (+1.33)
BASE-L3	22.45M	3	768	12	3072	67.31 (+2.21)	**66.03** (+0.67)	66.73 (+0.79)
LARGE-L4	52.49M	4	1024	16	4096	65.98 (+0.88)	64.65 (−0.71)	65.67 (−0.27)

Table 8: **Out-of-distribution head-to-head: SMALL-L2 vs. BASE-L3.** DAPFAM nDCG@100 (test split); LoCo macro-avg nDCG@10 (12 subtasks). BASE-L3 wins 11 of 12 cells. Deltas are computed on unrounded scores.
Bold = winner per cell.

Benchmark	GN, stride	small-l2	base-l3	Δ
DAPFAM	GTE-small,384	30.68	**31.10**	+0.42
DAPFAM	GTE-small,512	30.60	**31.69**	+1.10
DAPFAM	GTE-base,384	31.92	**32.72**	+0.80
DAPFAM	GTE-base,512	31.68	**32.68**	+1.00
DAPFAM	GTE-large,384	32.55	**33.06**	+0.51
DAPFAM	GTE-large,512	32.64	**33.10**	+0.45
LoCo	GTE-small,384	68.03	**68.87**	+0.84
LoCo	GTE-small,512	67.72	**68.92**	+1.20
LoCo	GTE-base,384	66.97	**67.74**	+0.77
LoCo	GTE-base,512	66.84	**67.72**	+0.88
LoCo	GTE-large,384	**70.97**	70.77	−0.20
LoCo	GTE-large,512	70.42	**70.49**	+0.07

Table 9: **Chunk-position ablation.** nDCG@10 on GOODWIKI-LONG test and LoCo (macro over 12 subtasks), nDCG@100 on DAPFAM test; deltas vs. the no-PE arm, computed on unrounded scores.

Chunk-position signal	GoodWiki	LoCo	DAPFAM
None (our design)	**64.52**	**65.25**	**30.46**
Learned absolute	64.45 (−0.08)	65.12 (−0.13)	29.88 (−0.58)
Sinusoidal	62.36 (−2.16)	59.48 (−5.77)	21.40 (−9.06)

with both terms non-negative. The first is an intrinsic property of the frozen GN, fixed at the value in (i). The second equals $I(X; E \mid R_\theta)$ and is the only term θ moves.

Proof.

(i) T and E are deterministic functions of X, so $I(X;T) = H(T)$ and $I(X;E) = H(E)$, both finite, and $E = g(T)$ gives $H(T,E) = H(T)$. Hence $I(X;T) - I(X;E) = H(T) - H(E) = H(T \mid E) \geq 0$. For countably valued T, $H(T \mid E) = 0$ iff T is almost surely a function of E, i.e. iff token sequences of positive probability receive distinct GN embeddings; collisions on the rest are harmless.

(ii) Both gaps are non-negative by the data-processing inequality along $X \to T \to E$ and $X \to E \to R_\theta$, and adding and subtracting $I(X;E)$ in $I(X;T) - I(X;R_\theta)$ gives the decomposition. Since R_θ is a function of E we have $I(X; R_\theta \mid E) = 0$, so expanding $I(X; E, R_\theta)$ by the chain rule in both orders leaves $I(X;E) - I(X;R_\theta) = I(X; E \mid R_\theta)$.

□

Proposition 2 (Sufficiency closes the trainable gap)**.** *Under the assumptions of Theorem 1, the trainable gap $I(X;E) - I(X;R_\theta)$ vanishes iff $X \perp\!\!\!\perp E \mid R_\theta$, that is, iff R_θ is a sufficient statistic of E for X, and in particular whenever h_θ is injective on the support of E.*

Proof. The gap equals $I(X; E \mid R_\theta)$ by Theorem 1(ii), which is zero exactly under that conditional independence. Injectivity on the support of E makes E almost surely a function of R_θ, so $0 \leq I(X; E \mid R_\theta) \leq H(E \mid R_\theta) = 0$. □

Proposition 3 (Contrastive recovery under InfoNCE)**.** *Let (E, E^+) be a positive pair under a stochastic augmentation $t \sim \mathcal{T}$ and let $\mathcal{L}_{\text{NCE}}(\theta)$ be the InfoNCE objective over batches of $B > 1$ pairs drawn i.i.d., each anchor scored against its own positive and $B - 1$ independent draws from the marginal of E^+, with critic $(e^+, r) \mapsto \text{sim}(h_\theta(e^+), r)/\tau$. Then*

$$I(E^+; R_\theta) \;\geq\; \log B \;-\; \mathcal{L}_{\text{NCE}}(\theta)$$

for every θ (Oord et al., 2018, Appendix A.1).

The bound applies at every θ, not only at the optimum, and tightens as the loss falls, but it certifies at most $\log B$ of mutual information.

Remark 4 (What the cosine objective gives). The released recipe admits no analogue of Proposition 3, since its per-pair terms are not normalised

over a candidate set and so support no likelihood-ratio bound on mutual information. It does bear on the condition of Proposition 2. Consider two documents with nonzero embeddings $e \neq e'$ that appear together as an $s{=}{-1}$ pair. Their cost $c_+ = \max\big(0, \cos(h_\theta(e), h_\theta(e'))\big)$ is at most 1, with equality iff the two embeddings point the same way, the collapse $h_\theta(e) = h_\theta(e')$ included, and strictly below 1 otherwise. The negative branch thus penalises, on the negative pairs seen in training, the collapses that violate the injectivity condition of Proposition 2. We claim only that direction of pressure, with no bound on $I(X; E \mid R_\theta)$ and no property of the minimisers.

Interpretation. Statement (i) prices the swap of token space for GN embeddings at $H(T \mid E)$, an *irreversible* loss that no choice of θ recovers once the GN is fixed. Statement (ii) leaves the remainder to training, and Proposition 2 names its target. REIGN cannot recover what the GN discarded, but *can* in principle preserve everything E carries about X, the formal counterpart of the empirical claim that REIGN's quality is upper-bounded by its GN (Limitations, "Coupling to the Guidance Network"). Our two objectives differ in what supports that "in principle". InfoNCE brings the quantitative handle of Proposition 3; the released cosine recipe brings the qualitative pressure of Remark 4 and the benchmark comparison of Appendix H, where it beats every InfoNCE arm we ran.

F Efficiency Analysis

Table 10 reports concrete training and inference FLOPs for representative long-document embedding models under a fixed 8K-token setting with 512-token chunks ($k{=}16$). For token-level Transformer models, we approximate the cost of a standard encoder layer as $\mathcal{O}(16nd^2 + 2n^2d)$, where n is the token sequence length, accounting for linear projections in the attention and feed-forward blocks as well as the quadratic self-attention term, following established practices in prior efficiency analyses (Vaswani et al., 2017; Kaplan et al., 2020). For Longformer-style sparse attention, the quadratic term is reduced to $\mathcal{O}(2nwd)$, where w denotes the local attention window size (Beltagy et al., 2020).

All FLOPs counts treat a multiply–add as two floating-point operations. Training FLOPs are computed as three times the forward-pass cost to account for forward and backward propagation, with additional overhead from the contrastive objective at batch size $B{=}48$; this B is a fixed analytic setting applied to all systems, not the training batch size of Appendix G. Unless otherwise stated, GTE-large, Longformer-large, and the full-attention 8K encoder assume hidden dimension $d{=}1024$ and $L{=}24$ encoder layers. REIGN pairs a frozen GTE Guidance Network with a $1024{\to}384$ projection and a lightweight two-layer (SMALL-L2) encoder over $k{=}16$ chunk-level embeddings. When Guidance Network embeddings are cached, REIGN avoids repeated token-level computation entirely, resulting in several orders-of-magnitude reductions in both training and inference FLOPs relative to token-level long-context baselines.

Measured cost. Table 11 complements the analytic model with measured end-to-end latency and peak memory on a single RTX 4090 (24 GB). We time 500 corpus documents and 100 queries from the GOODWIKI-LONG test split at batch 8, with one warm-up plus three timed repeats on an otherwise-idle GPU. Cached REIGN serves GN chunk embeddings from disk and answers queries in 0.40–0.52 ms (49–229× faster than with the GN run per query, i.e. the uncached rows); the one-time cache build costs 8.9 ms (GTE-small), 18.0 ms (GTE-base), and 53.3 ms (GTE-large) per document. Uncached REIGN performs identical GN work to the chunked baseline (verified at 167 chunks in both paths on the sampled documents), so we report parity; the sub-1.0 ratios reflect baseline batching, not less work. Two caveats remain. The cached peak-memory figures still include the GN weights resident on the device, which a corpus-only indexing deployment would not need, and Jina-v3 could not run at batch 8 on this card (measured at batch 4, peaking at 18.9 GB).

G GoodWiki-Long Training Protocol

Optimisation. REIGN is trained on GOODWIKI-LONG with the three-way cosine embedding objective of Section 4.2 (partial weight $\lambda{=}0.5$). Optimiser AdamW with lr$=10^{-5}$ and weight decay 10^{-4}; cosine-annealing schedule. Batch size is 18. Each anchor is paired with its rephrased positive (target 1), its two topical distractors as partials (target 0), and 17 in-batch negatives formed by shifting the positives (target -1), giving 360 pairs per optimisation step. Training runs 50 epochs with validation every 4 epochs; the released checkpoints are the best-validation snapshots (nDCG@10 on the `val` split). GN chunk embeddings are precomputed

Table 10: **Concrete FLOPs comparison for long-document embedding methods (per document)** under an 8K-token setting with 512-token chunks (k=16). Token-level baselines run self-attention over subword tokens; REIGN also runs full self-attention but over chunk-level embeddings (sequence length k=16 instead of n=512 or $8K$), and caches the GN forward pass. Multiply–add operations count as 2 FLOPs; training FLOPs assume 3× forward passes (forward + backward) plus the contrastive-objective overhead at batch size B=48.

Model	Attention Type	Representation	Training TFLOPs	Inference TFLOPs	Scaling
GTE (truncated, 512)	Full self-attn	Token-level	0.657	0.219	$\mathcal{O}(n^2)$
GTE (chunked, 16×512)	Full self-attn (repeated)	Token-level	10.514	3.505	$\mathcal{O}(kn^2)$
Longformer (full 8K)	Sparse windowed (w=512)	Token-level	10.509	3.503	$\mathcal{O}(nw)$
Full-attention encoder (8K)	Full self-attn	Token-level	19.785	6.595	$\mathcal{O}(n^2)$
REIGN (small-l2, *no cache*)	Full self-attn	Embedding-level	10.514	3.505	$\mathcal{O}(kn^2)$
REIGN (small-l2, cached GN)	Full self-attn	Embedding-level	**0.00027**	**0.000088**	$\boldsymbol{\mathcal{O}(k^2)}$

Table 11: **Measured inference cost** on GOODWIKI-LONG (500 documents, 100 queries), single RTX 4090 (24 GB), otherwise-idle GPU. Latency is end-to-end per query (encode + retrieve).
Uncached runs the Guidance Network per query; *cached* serves its chunk embeddings from disk.
Jina-v3 measured at batch 4 (batch 8 exceeds 24 GB); all other dense rows at batch 8.

System	Index (s)	ms/query	Peak GPU (GB)
Sparse lexical			
BM25	0.2	6.5	–
TF-IDF	0.1	1.1	–
Native long-context dense (truncate 8K)			
BGE-M3	37.6	430.3	9.80
Jina-Embeddings-v3	26.0	336.7	18.90
Stella-en-1.5B-v5	30.6	318.9	14.34
Nomic-Embed-v1.5	13.4	160.5	4.81
Bare GN (chunked mean-pool, 512)			
GTE-small (chunked)	4.1	23.0	0.23
GTE-base (chunked)	7.7	42.7	0.61
GTE-large (chunked)	21.6	121.4	1.56
BGE-base (chunked)	7.7	42.9	0.61
BGE-large (chunked)	21.6	121.4	1.56
REIGN, uncached GN (cold cache)			
REIGN + GTE-small	3.3	19.8	0.35
REIGN + GTE-base	6.6	39.7	0.76
REIGN + GTE-large	19.9	118.2	1.73
REIGN, cached GN embeddings (warm)			
REIGN + GTE-small	0.2	0.40	0.24
REIGN + GTE-base	0.2	0.47	0.55
REIGN + GTE-large	0.2	0.52	1.45

and cached. 16-bit mixed precision, seed 42.

Encoder configuration. The paper-default BASE-L3 REIGN encoder is a 3-layer Transformer over chunk-level embeddings (22.45M parameters; see Table 7 for the full configuration sweep). An internal linear projection adapts the GN's chunk-embedding dimensionality to the REIGN encoder hidden width when the two differ (e.g. 1024→768 for GTE-LARGE). Chunking uses a sliding window of K=512 tokens with stride S=K unless otherwise specified.

Data and evaluation. Training uses the GOODWIKI-LONG-SYNTHETIC `train` split (10,000 query-disjoint queries; positives enter the loss at full weight and distractors as λ-weighted partials, and the graded labels also provide the test-time ranking signal documented in Table 1). Reported test numbers (Table 2) are nDCG@10 on the 5,854-query `test` split with the full 53,562-document corpus; self-matches are removed. nDCG uses exponential graded gains ($2^{\mathrm{rel}}-1$), so the rephrased positive (score 2) carries gain 3 and each topical distractor (score 1) gain 1; the auxiliary precision/recall metrics in the released result files use linear gains $\mathrm{rel}/2$. We apply the same protocol and metric implementation to every system in Table 2 (sparse, native long-context dense, bare-GN, and REIGN). All runs use a single 24 GB consumer GPU.

H Training-Objective Ablation

Table 12 compares the three-way cosine recipe of Section 4.2 against InfoNCE variants under matched conditions (BASE-L3 on GTE-SMALL, chunk/stride 512, 20 epochs unless noted, seed 42, best-validation selection). The InfoNCE grid covers temperature $\tau \in \{0.07, 0.1\}$, distractors included as graded soft positives (α=0.5) or excluded (α=0), batch size 18 vs. 48, a warm start from the released checkpoint, and a 50-epoch run. The cosine recipe wins every benchmark against every InfoNCE variant; its 20-epoch replica (66.99) tracks the released 50-epoch checkpoint at the matched stride (67.09 @s512), so training length does not explain the gap. Within the InfoNCE family the graded distractor signal does not help. The α=0 arms beat their α=0.5 counterparts on all three benchmarks at τ=0.07 and on two of three at τ=0.1. Whether the pairwise cosine objective would also gain from excluding distractors is untested; the released recipe includes them as λ-weighted partials. The same ordering holds on short-text MTEB retrieval (ARGUANA/FIQA-2018 at 50.94/33.62 nDCG@10 for the 20-epoch

cosine arm, best of the arms measured on MTEB; the released checkpoint's MTEB numbers are in Table 6).

Table 12: **Training-objective ablation** (BASE-L3 on GTE-SMALL, chunk/stride 512, 20 epochs unless noted). nDCG@10 on GOODWIKI-LONG test and LoCo (macro over 12 subtasks), nDCG@100 on DAPFAM test. The reference row reports the released checkpoint's main-table values at their respective strides (GoodWiki @s384; 67.09 at the matched s512; LoCo and DAPFAM @s512).
α = distractor (soft-positive) weight; bs = batch size; **bold** = best ablation arm per column.
†Validation only degrades after the first validation epoch; the row is the epoch-1 best-validation snapshot.
‡Warm start from the released checkpoint; validation decays monotonically from the warm-start value, so the row is an epoch-1 snapshot close to the released model.

Objective	GoodWiki	LoCo	DAPFAM
Three-way cosine, bs18 (recipe)	**66.99**	**68.26**	**31.60**
Three-way cosine, bs48	65.42	66.77	31.11
InfoNCE τ=.07, α=0, bs48	65.91	65.94	30.71
InfoNCE τ=.07, α=0, bs48, 50 ep	66.37	66.20	30.49
InfoNCE τ=.07, α=.5, bs48	64.52	65.25	30.46
InfoNCE τ=.07, α=.5, bs48, warm‡	65.74	65.88	30.87
InfoNCE τ=.1, α=0, bs48	64.76	64.95	30.64
InfoNCE τ=.1, α=.5, bs48	64.25	64.98	30.19
InfoNCE τ=.07, α=.5, bs18†	60.16	63.15	29.48
Released ckpt (cosine, 50 ep)	*67.31*	*68.92*	*31.69*

Reproducibility note: InfoNCE needs a large negative pool. The InfoNCE path contrasts each anchor against the shifted in-batch positives only, so at batch 18 the per-step pool is 324 pairs, versus 360 pairs including partials on the cosine path. At τ=0.07 this pool is too small for stable training. Validation nDCG@10 peaks at the first validation epoch (0.73) and decays monotonically to 0.20 by epoch 19, so the batch-18 row above is an epoch-1 snapshot. Batch 48 (2,304 pairs) removes the collapse; the α=0 arm improves steadily (0.76 $\rightarrow$ 0.78), while the α=0.5 arm still decays after an early peak (0.76 $\rightarrow$ 0.61). The cosine objective shows no such sensitivity, and its validation improves monotonically at both batch sizes.

I DAPFAM Protocol Details

Dataset. DAPFAM (Ayaou et al., 2025) pairs 1,247 query patent families with a corpus of 45,336 candidate families, each represented by the concatenated text of its constituent patent documents. Relevance is binary, averaging 19.99 cited positives and 20 provided random non-cited negatives per query. Sampled whitespace-token length statistics give queries a median of 12,330 tokens (p75 21,204, p95 60,616, max 485,712) and corpus documents a median of 7,346 tokens (p75 12,262, p95 31,952, max 294,417). We use the FULLTEXT view throughout, which is the long-document setting REIGN targets.

Split construction. DAPFAM ships eval-only. To train REIGN on patents we built a query-disjoint train/val/test partition: (i) stratify the 1,247 query patents by quartiles of positive count; (ii) split randomly $\sim$70/15/15 with seed 42, query-disjoint (no query appears in more than one split); (iii) scope the original IN-/OUT-IPC partition to the held-out test set, so TEST_IN and TEST_OUT are positives-only subsets of TEST and the cross-domain breakdown is computed on identical queries across systems. The resulting counts are 872/188/187 unique queries (34,869/7,520/7,480 qrels rows) for train/val/test, with 180 and 138 queries in TEST_IN and TEST_OUT respectively.

Training protocol. REIGN is fine-tuned on DAPFAM via a standard query/positive/negative contrastive setup with InfoNCE and false-negative masking. Binary qrels disable the objective's partial branch (`partial-policy = ignore` in the released configuration). Negatives are DAPFAM's provided random `score`=0 families (`n-negatives-per-sample` = 4 plus in-batch negatives; using all 20 provided negatives exceeds 24 GB of memory at FullText sequence length). Warm-start runs initialise from the GOODWIKI-LONG-trained REIGN checkpoint; cold-start runs initialise the REIGN encoder from scratch with the same frozen GN. Optimiser AdamW, lr $\in \{10^{-5}, 5{\cdot}10^{-6}, 2{\cdot}10^{-6}, 10^{-6}\}$, weight decay $\in \{10^{-4}, 10^{-2}, 10^{-1}\}$, cosine schedule, temperature 0.07, 16-bit mixed precision, seed 42, validation every 3 epochs.

Evaluation protocol. The authoritative metric is nDCG@100 computed with a qrels-based evaluator over the full 45,336-document FullText corpus at top-k = 100; the in-training validation nDCG@10 is an in-batch proxy only and is *not* the reported number. Self-matches are removed before scoring: for typically $\sim$12–16 of the 187 test queries the query document itself appears in the corpus, and those corpus copies are dropped from the candidate pool. The same protocol is applied identically to zero-shot REIGN, fine-tuned REIGN, and every sparse or dense baseline; only the encoder varies.

Significance re-runs. Two deltas in Table 5 come from re-runs. Jina-v3 scores 32.98 against the main-table 32.97 (fp16 non-determinism) and Stella-1.5B 33.00 against 32.91 (batch 8 rather than the main-table batch 4); every other row matches Table 4 exactly, and holding Stella at its main-table 32.91 leaves the verdict unchanged (p=.843).