DISTILLING BIDIRECTIONAL EMBEDDING TEACHERS INTO STREAMING-COMPATIBLE CAUSAL STUDENTS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Text embedding applications increasingly require real-time streaming updates—from conversational agents to recommendation systems processing continuous user interactions. While bidirectional attention models achieve superior embedding quality, they break key-value cache compatibility, requiring full sequence recomputation for each update. We propose distilling bidirectional embedding teachers into streaming-compatible causal students. Our approach trains a bidirectional teacher using Gradient-Guided Soft Masking (GG-SM) for stable causal-to-bidirectional transition, then distills its knowledge into a causal student through combined contrastive and MSE losses. The distilled student achieves 68.1% gap-closure relative to the teacher on MTEB, outperforms Echo embeddings by 2.0 percentage points without the $2\times$ token overhead, and enables $4.1\times$ streaming speedup through KV-cache reuse. Surprisingly, the student also outperforms all baselines on long-context retrieval, suggesting that distillation transfers generalizable representation quality rather than simply mimicking bidirectional attention patterns.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Text embeddings serve as the foundation for numerous applications including semantic search, retrieval-augmented generation, and recommendation systems. Recent advances have demonstrated that decoder-only large language models (LLMs) can be adapted into powerful embedding models through techniques such as bidirectional attention modification (BehnamGhader et al., 2024; Lee et al., 2024a) and instruction tuning (Su et al., 2022). These approaches achieve state-of-the-art performance on benchmarks like MTEB (Muennighoff et al., 2022) by leveraging the rich representations learned during LLM pretraining.

However, many real-world applications require *streaming* embedding updates where text grows incrementally—chat sessions accumulate messages, user interaction histories extend over time, and documents undergo continuous editing. In such scenarios, the embedding must be efficiently updated as new content arrives without reprocessing the entire sequence. Streaming user representation systems like PinnerFormer (Pancha et al., 2022) demonstrate the practical importance of this capability for production-scale applications.

Bidirectional attention, while beneficial for embedding quality, fundamentally breaks streaming compatibility. Under causal attention, key-value (KV) caches enable $O(\Delta)$ incremental updates by reusing cached states for the prefix. Bidirectional attention invalidates this property: each new token changes the attention patterns for all positions, requiring $O(L + \Delta)$ full recomputation. Existing approaches to improve causal embeddings, such as Echo embeddings (Springer et al., 2024) which duplicate the input at inference time, incur $2\times$ token overhead and still require full recomputation for each update.

We propose a teacher-student distillation framework that transfers the quality benefits of bidirectional attention to a streaming-compatible causal student. A bidirectional teacher is first trained

[1] https://gitlab.com/fars-a/ggsm-causal-embedding-distill

using GG-SM progressive soft masking (Yuan et al., 2026) for stable causal-to-bidirectional transition. The causal student then learns from both contrastive signals (InfoNCE loss) and the teacher's embeddings (MSE distillation loss), achieving a balance between quality and streaming efficiency.

Our contributions are as follows:

- We introduce a teacher-student distillation framework that enables streaming-compatible text embeddings while preserving bidirectional embedding quality, combining InfoNCE contrastive loss with MSE distillation from a GG-SM-trained bidirectional teacher.
- The distilled student achieves 68.1% gap-closure relative to the bidirectional teacher on MTEB-slice (0.623 vs 0.645), outperforming Echo embeddings by 2.0 percentage points while enabling 4.1× mean latency speedup through KV-cache streaming.
- We analyze task-type variation in distillation effectiveness, finding strong transfer for classification (1.03×) and clustering (0.80×) but limited transfer for pair classification (0.30×), suggesting that embedding-level distillation primarily captures holistic semantic structure.

2 Related Work

Text Embeddings. The evolution of text embeddings has progressed from early encoder-based approaches to modern LLM-based methods. Sentence-BERT (Reimers & Gurevych, 2019) pioneered efficient sentence embeddings using siamese networks, while SimCSE (Gao et al., 2021) demonstrated that simple contrastive learning with dropout-based augmentation yields strong representations. E5 (Wang et al., 2022) scaled contrastive pre-training with weakly-supervised data, establishing new benchmarks on MTEB (Muennighoff et al., 2022). These encoder-based methods rely on bidirectional attention to capture full contextual information.

LLM-based Embeddings. Recent work has explored adapting decoder-only LLMs for embedding tasks. Bidirectional approaches such as LLM2Vec (BehnamGhader et al., 2024) and NV-Embed (Lee et al., 2024a) modify attention masks to enable full bidirectional context, achieving state-of-the-art performance. GRIT (Muennighoff et al., 2025) unifies generative and representational capabilities in a single model. However, these bidirectional modifications break KV-cache compatibility, preventing efficient streaming inference. Causal approaches like SGPT (Muennighoff, 2022) and Causal2Vec (Lin et al., 2025) maintain streaming compatibility but sacrifice embedding quality. Echo embeddings (Springer et al., 2024) provide an inference-time solution by duplicating input tokens, but incur 2× token overhead and still require full recomputation for updates.

Knowledge Distillation for Embeddings. Knowledge distillation has been successfully applied to compress embedding models. DistilBERT (Sanh et al., 2019) demonstrated that smaller models can retain most of a teacher's performance through distillation. Gecko (Lee et al., 2024b) distills LLM-generated synthetic data into compact embedding models. SimTDE (Xie et al., 2023) proposes simple transformer distillation for sentence embeddings. Our work differs by distilling across attention architectures—from bidirectional teacher to causal student—to enable streaming compatibility while preserving embedding quality.

Streaming Embeddings. Efficient streaming inference is critical for real-time applications. PinnerFormer (Pancha et al., 2022) demonstrates the importance of efficient user representation updates in recommendation systems. KV-Embedding (Tang & Yang, 2026) explores training-free approaches to leverage KV-cache for embeddings. Our distillation framework enables streaming-compatible embeddings that maintain quality comparable to bidirectional models while supporting efficient incremental updates through KV-cache reuse.

3 Method

We present a teacher-student distillation framework that enables streaming-compatible text embeddings while preserving the quality benefits of bidirectional attention. Figure 1 illustrates our approach.

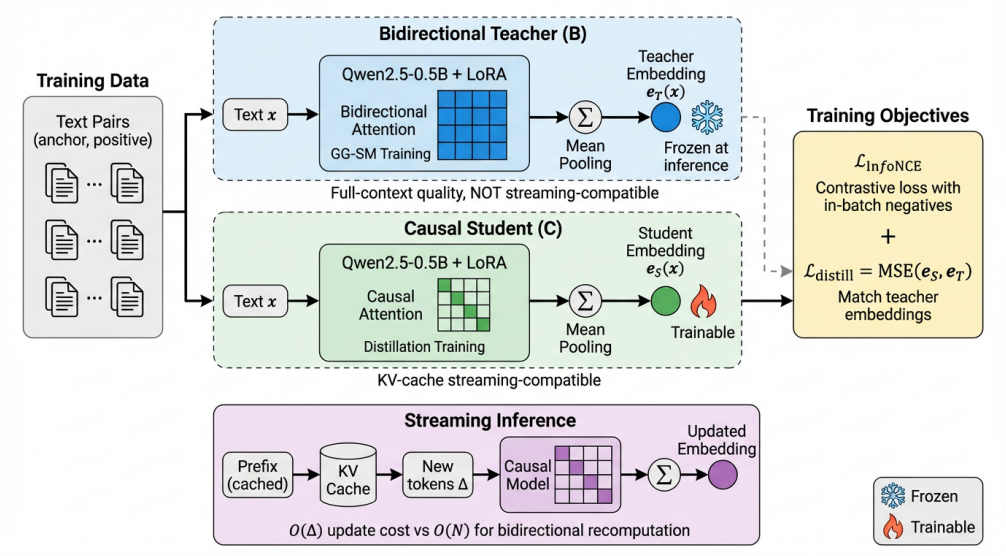

Figure 1: Overview of the teacher-student distillation framework. The bidirectional teacher (trained with GG-SM progressive soft masking) generates target embeddings, which guide the causal student through MSE distillation loss combined with contrastive InfoNCE loss. The student maintains causal attention for KV-cache compatibility, enabling efficient streaming inference.

3.1 Problem Formulation

Consider a streaming application where text grows incrementally as tokens are appended (e.g., chat sessions, user interaction histories, continuously edited documents). Let $x = (x_1, \ldots, x_L)$ denote the current sequence and Δx denote newly appended tokens. The goal is to efficiently update the embedding $e(x \oplus \Delta x)$ without reprocessing the entire sequence.

Decoder-only LLMs with causal attention support efficient incremental inference through key-value (KV) caching. Under causal attention, the hidden state h_i of token i depends only on tokens $1, \ldots, i$. When new tokens arrive, the cached key-value pairs for the prefix remain valid, and only the new tokens require forward computation. This enables $O(\Delta)$ update complexity rather than $O(L + \Delta)$ full recomputation.

However, bidirectional attention—which allows each token to attend to all other tokens—breaks this property. Adding new tokens changes the attention patterns for all positions, invalidating the cached states and requiring full sequence recomputation. While bidirectional models like LLM2Vec (BehnamGhader et al., 2024) and NV-Embed (Lee et al., 2024a) achieve superior embedding quality by leveraging full contextual information, they sacrifice streaming efficiency.

3.2 Teacher Training with GG-SM

We train a bidirectional teacher using Gradient-Guided Soft Masking (GG-SM) (Yuan et al., 2026), which provides a stable transition from causal to bidirectional attention. The key insight is that abruptly switching from causal to bidirectional masking can destabilize training, as the model must suddenly adapt to a fundamentally different attention pattern.

GG-SM addresses this through a two-phase schedule. During the warmup phase ($t < T_{\text{warm}}$), the soft attention mask M_{soft} is defined as:

$$M_{\text{soft},ij}(t) = \begin{cases} 0 & \text{if } j \leq i \\ \log \sigma(\|\nabla_{h_j} \mathcal{L}\|) & \text{if } j > i \end{cases} \tag{1}$$

where σ is the sigmoid function and $\nabla_{h_j} \mathcal{L}$ is the gradient of the loss with respect to hidden state h_j. This gradient-guided weighting allows the model to gradually attend to future tokens based on their relevance to the training objective.

After warmup ($t \geq T_{\text{warm}}$), the weights are frozen and linearly interpolated toward full bidirectionality:

$$w_{ij}(t) = (1 - \alpha_t) \cdot \sigma(\|\nabla_{h_j} \mathcal{L}_{\text{warm}}\|) + \alpha_t \tag{2}$$

where $\alpha_t = (t - T_{\text{warm}})/(T_{\text{total}} - T_{\text{warm}})$. At the end of training, the teacher has full bidirectional attention and can leverage complete contextual information for embedding extraction.

3.3 Student Training with Distillation

The student maintains standard causal attention to preserve KV-cache compatibility. We train it with a combined objective that balances contrastive learning with knowledge distillation from the teacher.

The contrastive loss uses InfoNCE (van den Oord et al., 2018) with in-batch negatives:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(e_S(x), e_S(x^+))/\tau)}{\sum_{x' \in \mathcal{B}} \exp(\text{sim}(e_S(x), e_S(x'))/\tau)} \tag{3}$$

where e_S is the student embedding function, x^+ is a positive pair, $\mathcal{B}$ is the batch, and τ is the temperature.

The distillation loss minimizes the MSE between student and teacher embeddings:

$$\mathcal{L}_{\text{distill}} = \|e_S(x) - e_T(x)\|_2^2 \tag{4}$$

where e_T is the frozen teacher embedding. Both embeddings are L2-normalized before computing the loss.

The final training objective combines both losses:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{distill}} \tag{5}$$

with $\lambda = 0.5$ balancing the two objectives. Teacher embeddings are precomputed and cached to avoid repeated forward passes during student training.

3.4 Streaming Inference

At inference time, the student uses standard causal attention with KV-cache for efficient streaming updates. For embedding extraction, we use mean pooling over content token hidden states:

$$e(x) = \text{norm}\left(\frac{1}{|S|} \sum_{i \in S} h_i\right) \tag{6}$$

where S is the set of content token indices (excluding special tokens) and $\text{norm}(\cdot)$ denotes L2 normalization.

Mean pooling is naturally compatible with streaming updates. We maintain a running sum $s = \sum_{i \in S} h_i$ and count $|S|$. When new tokens arrive, we forward only the new tokens using the KV-cache, add their hidden states to s, update $|S|$, and output $\text{norm}(s/|S|)$. This requires no recomputation of the prefix, achieving $O(\Delta)$ complexity per update.

4 Experiments

4.1 Experimental Setup

We evaluate our distillation framework using Qwen2.5-0.5B-Instruct as the base model, fine-tuned with LoRA (Hu et al., 2021) (rank $r = 16$, $\alpha = 16$) targeting query, key, value, and output projection modules. All models are trained on the sentence-transformers/all-nli dataset using entailment pairs as positives, with sequence length 512, global batch size 256, and bfloat16 mixed precision.

We compare four conditions: (A) **Causal Baseline**—standard causal attention with InfoNCE loss only; (A+Echo) **Echo Embeddings** (Springer et al., 2024)—inference-time input duplication applied to Condition A; (B) **GG-SM Teacher**—bidirectional attention via progressive soft masking,

Table 1: Main results on MTEB-slice benchmark (6 datasets). Best in **bold**, second-best underlined. The distilled student (C) achieves 68.1% gap-closure relative to the bidirectional teacher (B), outperforming Echo embeddings while maintaining streaming compatibility.

Method	ArguAna	SciFact	STSBench	AmazonCF	SprintDQ	RedditClust	Mean
Causal (A)	0.466	0.384	0.747	0.679	0.793	0.389	0.576
Echo (A+Echo)	0.382	0.467	0.773	0.664	**0.892**	0.439	0.603
Teacher (B)	0.458	**0.546**	**0.790**	0.719	0.941	0.419	**0.645**
Student (C)	**0.508**	0.482	0.780	**0.721**	0.837	**0.413**	0.623

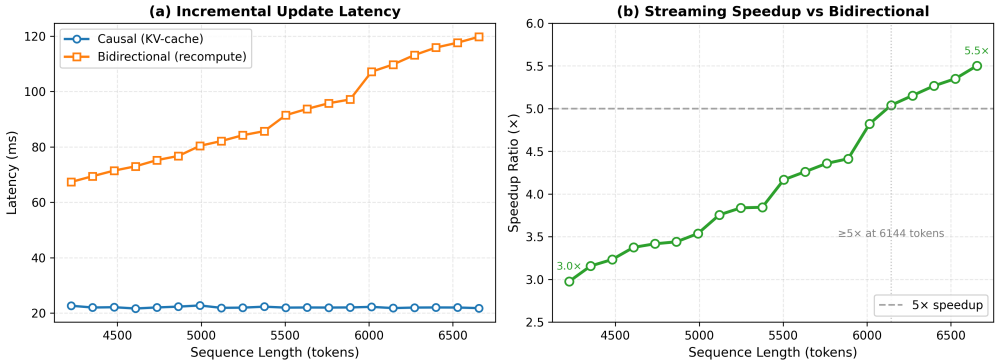

Figure 2: Streaming inference latency comparison. (a) Incremental update latency: causal attention with KV-cache maintains constant ∼22ms latency while bidirectional recomputation grows linearly (67→120ms). (b) Speedup ratio increases from 3.0× to 5.5× as sequence length grows from 4.2k to 6.6k tokens.

serving as the upper bound; and (C) **Distilled Student**—our method, causal attention trained with InfoNCE + MSE distillation from the teacher.

For evaluation, we use a 6-dataset slice of MTEB (Muennighoff et al., 2022) spanning diverse task types: ArguAna and SciFact (Retrieval, NDCG@10), STSBenchmark (STS, Spearman correlation), AmazonCounterfactualClassification (Classification, accuracy), SprintDuplicateQuestions (PairClassification, average precision), and RedditClustering (Clustering, V-measure). We additionally evaluate on LoCoV1 (Thakur et al., 2021) for long-context retrieval (NDCG@10). All experiments use 2 random seeds with mean results reported.

4.2 Main Results

Table 1 presents the main results on MTEB-slice. The distilled student achieves a mean score of 0.623, representing a 68.1% gap-closure ratio relative to the teacher: $(0.623 - 0.576)/(0.645 - 0.576) = 0.681$. This demonstrates that distillation successfully transfers most of the bidirectional teacher's quality gains to a streaming-compatible causal model.

The student outperforms the Echo embedding baseline by 2.0 percentage points (0.623 vs 0.603) while avoiding Echo's 2× token overhead and maintaining true streaming capability. Notably, the student exceeds the teacher on ArguAna (0.508 vs 0.458) and AmazonCounterfactualClassification (0.721 vs 0.719), suggesting that the distillation process may provide regularization benefits for certain task types.

4.3 Streaming Efficiency

Figure 2 demonstrates the streaming efficiency advantage of the causal student. We measure incremental embedding update latency starting from a 4096-token prefix and appending 128-token chunks for 20 updates. The causal student with KV-cache maintains constant latency of approximately 22ms per update regardless of total sequence length, while the bidirectional teacher requires full recomputation with latency growing linearly from 67ms to 120ms.

Table 2: Long-context retrieval results on LoCoV1 benchmark (NDCG@10). The distilled student outperforms all baselines including the bidirectional teacher, which surprisingly underperforms the causal baseline on long documents.

Method	LoCoV1 NDCG@10
Causal (A)	0.213
Echo (A+Echo)	0.227
Teacher (B)	0.212
Student (C)	**0.284**

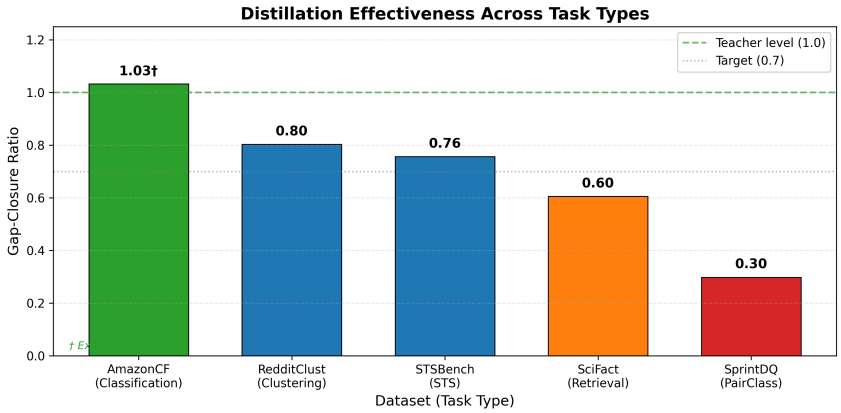

Figure 3: Gap-closure ratio across task types. The distilled student achieves strong transfer for classification (1.03, exceeding teacher) and clustering (0.80), moderate transfer for STS (0.76) and retrieval (0.60), but limited transfer for pair classification (0.30). Dashed lines indicate teacher level (1.0) and target threshold (0.7).

The mean speedup across all updates is 4.1×, reaching 5.0× at sequence lengths ≥6144 tokens. For longer sequences typical in production streaming applications (8K–16K+ tokens), the speedup advantage would be substantially larger. A prefix-fidelity check confirms that incremental updates produce numerically identical embeddings to full recomputation (maximum cosine distance 1.19×10^{-6}), validating the correctness of the streaming protocol.

4.4 Long-Context Evaluation

Table 2 presents results on LoCoV1 long-context retrieval. Surprisingly, the distilled student achieves the best performance (0.284 NDCG@10), outperforming all baselines including the bidirectional teacher by 7.1 percentage points over the causal baseline. The teacher actually underperforms the causal baseline on long documents (0.212 vs 0.213), likely due to position bias artifacts from training on short sequences (512 tokens) that do not transfer well to longer contexts.

This result suggests that distillation transfers generalizable representation quality rather than simply mimicking the teacher's bidirectional mechanism. The causal student's architecture is more robust to length extrapolation, and the distillation process appears to regularize against position-dependent artifacts present in the teacher.

4.5 Task-Type Analysis

Figure 3 analyzes gap-closure ratios across task types. Classification (1.03) and clustering (0.80) show strong transfer, with classification actually exceeding the teacher. STS (0.76) and retrieval (0.60) show moderate transfer. However, pair classification (0.30) shows limited transfer, falling well below the 0.7 target threshold.

This variation suggests that distillation is most effective for tasks requiring holistic semantic understanding (classification, clustering), where the global embedding captures sufficient information. Tasks requiring fine-grained pairwise comparison (pair classification) may depend more heavily on the bidirectional attention mechanism itself, which cannot be fully captured through embedding-level distillation alone.

5 Conclusion

We presented a teacher-student distillation framework that enables streaming-compatible text embeddings while preserving most of the quality benefits of bidirectional attention. Our distilled causal student achieves 68.1% gap-closure relative to the bidirectional teacher, outperforms Echo embeddings by 2.0 percentage points, and enables 4.1× streaming speedup through KV-cache reuse. Surprisingly, the student also outperforms all baselines on long-context retrieval, suggesting that distillation transfers generalizable representation quality.

Our approach has limitations: pair classification shows limited transfer (0.30 gap-closure), and we evaluated only a single model scale (0.5B parameters) with one training dataset. Future work could explore larger models, multi-teacher distillation, and task-specific distillation strategies to improve transfer for fine-grained comparison tasks.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *ArXiv*, abs/2404.05961, 2024.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, M. Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *ArXiv*, abs/2405.17428, 2024a.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernández Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha R. Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models. *ArXiv*, abs/2403.20327, 2024b.

Ailiang Lin, Zhuoyun Li, and Kotaro Funakoshi. Causal2vec: Improving decoder-only llms as versatile embedding models. *ArXiv*, abs/2507.23386, 2025.

Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *ArXiv*, abs/2202.08904, 2022.

Niklas Muennighoff, Nouamane Tazi, L. Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. pp. 2006–2029, 2022.

Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Nikil Pancha, Andrew Zhai, J. Leskovec, and Charles R. Rosenberg. Pinnerformer: Sequence modeling for user representation at pinterest. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. *ArXiv*, abs/2402.15449, 2024.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv*, abs/2212.09741, 2022.

Yixuan Tang and Yi Yang. Kv-embedding: Training-free text embedding via internal kv re-routing in decoder-only llms. *ArXiv*, abs/2601.01046, 2026.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663, 2021.

Aäron van den Oord, Yazhe Li, and O. Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533, 2022.

Jian Xie, Xin He, Jiyang Wang, Zimeng Qiu, Ali Kebarighotbi, and Farhad Ghassemi. Simtde: Simple transformer distillation for sentence embeddings. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Jiahao Yuan, Yike Xu, Jinyong Wen, Baokun Wang, Yang Chen, Xiaotong Lin, Wuliang Huang, Ziyi Gao, Xing Fu, Yu Cheng, and Weiqiang Wang. How do decoder-only llms perceive users? rethinking attention masking for user representation learning. 2026.