Causal2Vec: Improving Decoder-only LLMs as Embedding Models through a Contextual Token

Ailiang Lin[1*], **Zhuoyun Li**[2*], **Yusong Wang**[1], **Kotaro Funakoshi**[1], **Manabu Okumura**[1]
[1]Institute of Science Tokyo [2]Tencent
{linailiang, wangyi, funakoshi, oku}@lr.first.iir.isct.ac.jp
earyli@tencent.com

Abstract

Decoder-only large language models (LLMs) have been increasingly adopted to build embedding models for diverse tasks. To overcome the inherent limitations of causal attention in representation learning, many existing methods modify the attention mechanism to be bidirectional, potentially undermining LLMs' ability to extract semantic information acquired during pre-training. Meanwhile, leading unidirectional approaches often rely on extra input text to generate contextualized embeddings, inevitably increasing computational costs. In this work, we propose Causal2Vec, a general-purpose embedding model tailored to enhance the performance of decoder-only LLMs without altering their original architectures or introducing significant computational overhead. Specifically, we first employ a lightweight BERT-style model to pre-encode the input text into a single Contextual token, which is then prepended to the LLM's input sequence, allowing each token to capture contextualized information even without attending to future tokens. Furthermore, to mitigate the recency bias introduced by last-token pooling, we concatenate the last hidden states of Contextual and EOS tokens as the final text embedding. In practice, Causal2Vec achieves a new state-of-the-art performance on the MTEB benchmark among models trained solely on publicly available retrieval datasets.

1 Introduction

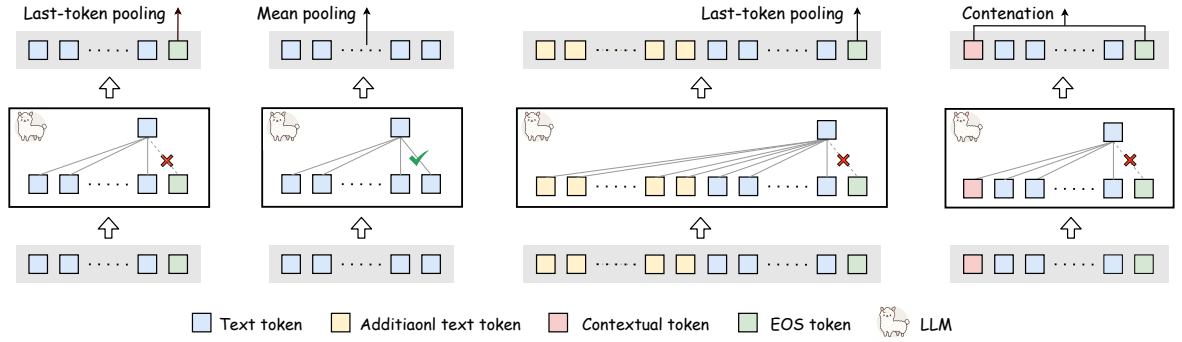
Text embedding models encode natural language text into dense vector representations that capture contextual semantic information (Conneau et al., 2017), enabling a wide range of downstream natural language processing (NLP) tasks, such as information retrieval, semantic textual similarity, and question answering (Xiong et al., 2021; Muennighoff et al., 2023; Jiang et al., 2026). Moreover, embedding-based retrievers play a crucial role in enhancing the capabilities of large language model (LLM)-based Retrieval-Augmented Generation (RAG) systems (Liu et al., 2024b; Zhang et al., 2026a,b). For many years, pre-trained language models based on encoder-only or encoder-decoder Transformer architectures, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020), have been the dominant paradigm for building text embedding models (Reimers and Gurevych, 2019; Gao et al., 2021; Ni et al., 2022; Wang et al., 2022).

With recent advances in LLMs, considerable efforts have focused on transforming decoder-only architectures into text embedding models. However, the use of causal attention in Transformer decoders leads to incomplete information encoding for each token except the last one (Figure 1-(a)), significantly limiting the model's representational capacity. To address this issue, many LLM-based text embedding methods (Muennighoff et al., 2024; BehnamGhader et al., 2024; Lee et al., 2025a; Pan et al., 2025) achieve bidirectional attention by removing the causal attention mask, enabling each token to access the entire sequence and thus generating rich contextualized representations, as illustrated in Figure 1-(b). Despite notable progress, our findings in Figure 3 highlight that modifying the original attention mechanisms of LLMs may be suboptimal for embedding tasks, as it leads to a pre-train/fine-tune attention mismatch, potentially compromising the model's ability to extract semantic information acquired during pre-training.

In contrast, most leading causal attention-based methods (Springer et al., 2025; Li et al., 2025) compensate for missing contextual information in the original sequence by introducing additional input text (Figure 1-(c)), which inevitably increases computational costs. Moreover, to derive text embeddings from the output hidden states of LLMs, there are two mainstream strategies: mean pooling and last-token pooling. Prior studies (Zhang et al.,

[*] Equal Contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22769–22788



(a) Causal Attention (b) Bidirectional Attention (c) Using additional input text (d) Ours

Figure 1: Comparison of different methods to overcome the representational bottleneck of decoder-only LLMs (Fig.a). Unlike existing methods that either modify attentions from causal to bidirectional (Fig.b) or utilize extra input text (Fig.c), we prepend a *Contextual token* to the LLM's input sequence (Fig.d). This allows each token to access contextualized information without attending to future tokens.

2023a; Li et al., 2025) show that, for embedding models based on causal attention, (weighted) mean pooling is less effective than last-token pooling, since autoregressive modeling prevents earlier tokens from attending to future tokens, leading to biased aggregated representations. As a result, last-token pooling has been more commonly adopted in unidirectional models. However, since last-token pooling typically relies on the final hidden state of the end-of-sequence (EOS) token, it can be highly sensitive to noisy information near the end of input and thus prone to recency bias (Springer et al., 2025; Lee et al., 2025a), hindering the model's ability to learn robust representations.

In this work, we propose **Causal2Vec**, a simple yet powerful causal attention-based embedding model that significantly enhances the text encoding capabilities of decoder-only LLMs, while circumventing the need to modify their original architectures or introducing significant computational overhead. Specifically, to address the representational bottleneck inherent in the causal attention mechanism while preserving the LLMs' ability to extract semantic information learned during pre-training, we first employ a lightweight, off-the-shelf bidirectional encoder to distill the contextual content of the input text into a single *Contextual token*, which is then aligned to the dimensionality of LLM's word embedding space via a trainable MLP layer. As shown in Figure 1-(d), by prepending this token to LLM's input sequence, we enable each token to access contextualized information even under the constraints of causal masks, without switching to bidirectional attention or utilizing extra input text. Moreover, we concatenate the last hidden states of Contextual and EOS tokens as the final text embedding, effectively mitigating the recency bias introduced by last-token pooling and encouraging LLMs to better leverage the contextualized information encoded in the Contextual token.

We conduct comprehensive experiments on the MTEB benchmark (Muennighoff et al., 2023) by integrating Causal2Vec into four decoder-only LLMs with parameter sizes ranging from 1.3B to 7B (S-LLaMA-1.3B, Qwen2.5-1.5B, LLaMA-2-7B, and Mistral-7B). Evaluation across 56 datasets spanning 7 tasks demonstrates that our method achieves new state-of-the-art results among models trained solely on publicly available retrieval data. Furthermore, we present extensive ablations and analyses to validate the effectiveness and necessity of the proposed mechanism. Overall, this work highlights the inherent potential of LLMs' original causal attention in generating high-quality contextualized text embeddings. The main contributions of this work are summarized below:

- We introduce Causal2Vec, a simple yet powerful approach that enhances the representational capacity of LLMs without converting to bidirectional attention or requiring extra input.
- To mitigate the representational bottlenecks of causal attention, we introduce the *Contextual token* and prepend it to LLM's input sequence, allowing each token to access contextual information without attending to future ones.
- To alleviate the recency bias introduced by last-token pooling, we concatenate the last hidden states of Contextual and EOS tokens as the final text embedding.

- Causal2Vec achieves state-of-the-art results on MTEB among models trained solely on publicly available retrieval datasets.

2 Related Work

Bidirectional Text Embedding Models. Over the past few years, embedding methods based on pre-trained language models with bidirectional attention, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020), have dominated text embedding tasks. Early notable approaches, including SimCSE (Gao et al., 2021) and Sentence-T5 (Ni et al., 2022), are pre-trained with a masked language modeling objective and fine-tuned in a contrastive manner with natural language inference (NLI) datasets. Later work like E5 (Wang et al., 2022) and GTE (Li et al., 2023b) further improves embedding performance through weakly supervised contrastive training on curated text pair datasets. More recent methods (Asai et al., 2023; Su et al., 2023; Wang et al., 2024b) have shifted toward developing general-purpose embedding models through task instructions, demonstrating strong generalization to unseen tasks.

Decoder-only LLM-based Text Embedding Models. Luo et al. (2024) show that larger models with extensive pre-training consistently improve performance in dense retrieval. Weller et al. (2025) further highlight that while bidirectional encoders excel at classification and retrieval tasks, the lack of large-scale encoder-only models in practice suggests that embedding models based on decoder-only LLMs will likely surpass all other options. Nevertheless, decoder-only LLM-based embedding methods still suffer from the inherent architectural drawbacks: causal attention prevents each token from interacting with subsequent tokens, hindering the model's ability to produce contextualized representations. To address this issue, BeLLM (Li and Li, 2024) transforms LLM's attention from unidirectional to bidirectional by removing the causal mask at specific layers. GRITLM (Muennighoff et al., 2024) and LLM2Vec (BehnamGhader et al., 2024) further enable fully bidirectional attention in LLMs. Building upon this attention modification, NV-Embed (Lee et al., 2025a) introduces a latent attention layer over LLM's output hidden states to generate higher-quality representations. Moreover, MGH (Pan et al., 2025) proposes a novel pooling method that dynamically aggregates the output sequence to acquire a more accurate text embedding.

Notably, these bidirectional LLM-based methods involve modifications to the model architecture, limiting their compatibility with diverse LLM backbones. In contrast, some studies preserve the original causal attention while attempting to alleviate its limitations. Wang et al. (2024a) employ proprietary LLMs to construct diverse synthetic data and fint-tune open-source decoder-only LLMs through standard contrastive learning, achieving competitive performance. ECHO (Springer et al., 2025) repeats the input twice in the autoregressive modeling paradigm, allowing the text embedding extracted from the repeated tokens to capture contextualized content. Similarly, TP (Fu et al., 2025) prepends the previous layer's last-token to the next layer's input sequence, enabling tokens to attend to the full sentence information. PromptEOL (Jiang et al., 2024) and bge-en-icl (Li et al., 2025) enhance text embeddings by leveraging the in-context learning (ICL) capabilities of LLMs, augmenting the original input with task-specific examples to provide contextual information. Furthermore, Anchor (Su et al., 2025) introduces a bidirectional reconstruction training stage before contrastive learning to enrich the semantics of the final embedding.

3 Method

Figure 2 illustrates an overall pipeline of the proposed Causal2Vec. Given an input text, we first use a lightweight bidirectional encoder to generate the Contextual token, which is then added to LLM's input sequence for causal attention computation. The final text embedding is derived by concatenating the last hidden states of the Contextual and EOS tokens. We elaborate on the Contextual token and representation method in the following sections.

3.1 Preserve the Causal Attention

The remarkable capacity of LLMs for human language understanding and generation is acquired through training on massive amounts of text data (Ouyang et al., 2022; Wei et al., 2022), showcasing their effectiveness in encoding semantic information. However, the inherent causal attention in LLMs prevents earlier tokens from accessing information about future tokens, thus hindering the model's representational capability.

To mitigate this limitation, many LLM-based embedding models switch from unidirectional attention to bidirectional by removing the causal attention mask. While bidirectional attention facil-

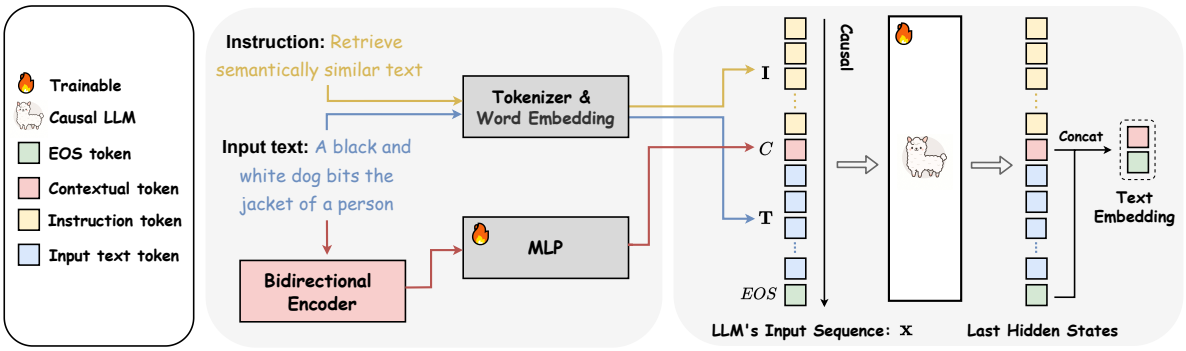


Figure 2: Overview of our Causal2Vec method.

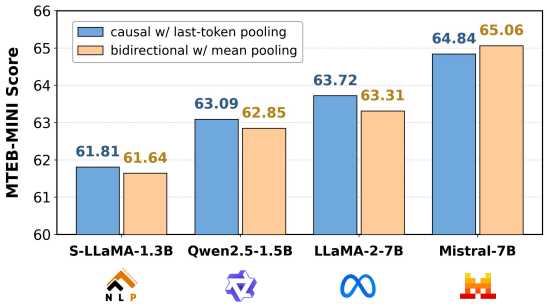


Figure 3: Average MTEB-MINI score (30 datasets) of causal vs. bidirectional attention across different LLM backbones. All results are obtained after contrastive learning on publicly available retrieval datasets.

itates effective information flow across the entire sentence, unlike models trained from scratch under bidirectional attention, this modification introduces an attention mismatch between pre-training and fine-tuning, potentially compromising the LLMs' ability to extract semantic information acquired during pre-training. As shown in Figure 3, altering the attention mechanism degrades embedding performance across most LLMs—except for Mistral-7B (Jiang et al., 2023), likely due to its non-standard pre-training methodology (Springer et al., 2025; BehnamGhader et al., 2024). Consequently, most existing bidirectional methods (Muennighoff et al., 2024; Lee et al., 2025a; Pan et al., 2025) rely on Mistral-7B as the foundation model, limiting their generalizability to diverse LLM architectures. Furthermore, Li et al. (2025) show that converting Mistral-7B to bidirectional harms its in-context learning capabilities for text embedding tasks.

These findings suggest that many LLMs do not benefit from attention modification, making it neither a robust nor a general solution for transforming decoder-only LLMs into text embedding models. Consequently, we preserve the original causal attention and instead explore alternative strategies to overcome its representational bottleneck. Please refer to Appendix C.6 for more comparisons.

3.2 The Contextual Token

To fully unlock the potential of decoder-only LLMs in generating text embeddings, it is essential to address the limitations of causal attention while preserving LLMs' ability to extract well-learned semantic information. To this end, we first introduce a lightweight BERT-style model that encodes the input text into a k-dimensional dense vector representation $h \in \mathbb{R}^{1\times k}$, termed the "*Contextual Token*". Specifically, this token is generated by applying mean pooling over the last hidden state of the additional bidirectional model, capturing contextualized information about the entire input. Furthermore, to bridge the gap between the BERT-style model and LLMs, we employ a simple MLP layer to match the dimensionality of the Contextual token with LLM's word embedding space, and then encourage the LLMs to understand the sentence information encoded in this token through contrastive learning. Motivated by Liu et al. (2024a), the MLP layer consists of two linear transformations with a GELU activation σ, which can be formulated as:

$$C = \sigma(h\mathbf{W}_1^\top)\mathbf{W}_2^\top, \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times k}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times d}$ are trainable projection matrices, and $C \in \mathbb{R}^{1\times d}$ denotes the language embedding of the Contextual token, which shares the same dimensionality d as the word embedding space of the LLMs. Moreover, to leverage LLMs' instruction-follow capabilities for producing general-purpose text embeddings, we use task-specific instructions for both training and evaluation, following Wang et al. (2024a); BehnamGhader et al. (2024); Springer et al. (2025). Consequently, by prepending the instruction and Contextual tokens and appending the EOS token to

the input sequence, the resulting sequence fed into the LLMs can be constructed as:

$$\mathbf{x} = [\mathbf{I}; C; \mathbf{T}; EOS] \in \mathbb{R}^{(l+n+2)\times d}, \quad (2)$$

where $[\cdot;\cdot]$ denotes the vertical concatenation operation, $\mathbf{I} \in \mathbb{R}^{l\times d}$ and $\mathbf{T} \in \mathbb{R}^{n\times d}$ represent the word embeddings of the task-specific instruction and input text, respectively. In this way, each token following the Contextual token C can capture contextualized information even without attending to future tokens. More importantly, the use of the Contextual token requires no modifications to the model architecture, which not only preserves LLMs' ability to extract semantic information learned during pre-training, but also enables seamless integration across different LLMs.

3.3 Representation Method

As the most widely adopted representation method for unidirectional models (Wang et al., 2024a; Li et al., 2025), last-token pooling typically utilizes the final hidden state of the EOS token as text embeddings, since only the last token captures information from the entire input. However, recent studies (Springer et al., 2025; Lee et al., 2025a) indicate that the EOS token embedding depends heavily on tokens near the end of the sequence, leading to potential semantic bias in long-text scenarios.

To address this issue, we introduce a simple yet effective representation method tailored for our embedding framework. Specifically, we concatenate the hidden states of the Contextual and EOS tokens from the LLM's last layer to generate text embeddings. Unlike last-token pooling, the Contextual token is not affected by tokens near the end of the sequence, thus effectively mitigating recency bias. In addition, since the Contextual token has already captured the semantic content of the input text, the concatenation of two context-aware tokens yields a vector representation with richer contextualized information. Furthermore, this approach enables explicit supervision of the Contextual token during training, thereby helping LLMs better understand the semantic information encoded in this added token. The proposed representation method is optimized through supervised contrastive learning with the standard InfoNCE loss (Izacard et al., 2021), which can be formulated as:

$$\mathcal{L} = -\log \frac{\exp(f(q,p^+)/\tau)}{\exp(f(q,p^+)/\tau) + \sum_{j=1}^{N} \exp(f(q,p_j^-)/\tau)}, \quad (3)$$

where $f(q,p^+)$ represents the scoring function that computes cosine similarity between the query-passage pair embeddings from retrieval datasets, p^- denotes both in-batch and hard negative examples, and τ is a temperature hyperparameter fixed at 0.05 in all experiments.

4 Experiments

4.1 Datasets

Training Datasets. For training, we follow the mainstream practices (BehnamGhader et al., 2024; Su et al., 2025; Pan et al., 2025; Li et al., 2025), utilizing the same collection of publicly available retrieval datasets curated by Springer et al. (2025), which consists of approximately 1.5 million samples. More details about the composition of the training datasets can be found in Appendix A.3.

Notably, many industry-developed embedding models, such as Qwen3 Embedding (Zhang et al., 2025), achieve strong performance by leveraging large amounts of non-public or proprietary synthetic data across various embedding tasks. To ensure fairness and consistency in academic comparisons, we evaluate only models trained on public retrieval datasets, enabling the verification of models' generalization capability to unseen non-retrieval tasks, which serves as a critical criterion for defining a general-purpose embedding model.

Evaluation Benchmark. For evaluation, we use the English subset of the Massive Text Embeddings Benchmark (MTEB) (Muennighoff et al., 2023), which comprises 56 datasets spanning 7 embedding task categories. Since MTEB is a large-scale benchmark containing over 35 million test samples, the full evaluation of a 7B parameter model is GPU resource intensive, requiring over 200 A100 80GB GPU hours. To speed up the evaluation, we curate a representative subset of MTEB, termed MTEB-MINI, for ablation studies and analysis. Specifically, MTEB-MINI consists of 30 datasets covering all task categories in MTEB. Details of the MTEB-MINI composition can be found in Appendix B.1.

4.2 Experimental Details

For the base model, we integrate Causal2Vec with four decoder-only LLMs ranging from 1.3B to 7B parameters: Sheared-LLaMA-1.3B (S-LLaMA-1.3B), Qwen2.5-1.5B-Instruct (Qwen2.5-1.5B), Llama-2-7B-chat (LLaMA-2-7B), and Mistral-7B-Instruct-v0.2 (Mistral-7B). Regarding the off-the-shelf bidirectional encoder, we adopt E5-base-

Task (# of datasets) Metric	Retr. (15) nDCG@10	Rerank. (4) MAP	Clust. (11) V-Meas.	PairClass. (3) AP	Class. (12) Acc.	STS (10) Spear.	Summ. (1) Spear.	Avg (56)
S-LLaMA-1.3B								
LLM2Vec (BehnamGhader et al., 2024)	51.44	55.38	43.57	86.20	72.21	83.58	30.01	61.85
ECHO (Springer et al., 2025)	-	-	-	-	-	-	-	62.01
Causal2Vec	52.69	56.54	44.35	86.18	72.94	83.76	31.45	62.63
Qwen2.5-1.5B								
Anchor (Su et al., 2025)	53.62	57.63	43.19	85.77	74.51	82.74	31.61	62.86
Causal2Vec	54.57	58.27	46.30	86.73	73.68	84.63	30.99	63.97
LLaMA-2-7B								
LLM2Vec (BehnamGhader et al., 2024)	54.60	57.38	45.24	88.03	76.33	83.73	28.49	64.14
Causal2Vec	55.28	58.18	47.23	87.85	75.95	84.90	31.09	64.94
Mistral-7B								
$E5_{Mistral\text{-}7b}$ (Wang et al., 2024a)	52.78	60.38	47.78	88.47	76.80	83.77	**31.90**	64.56
ECHO (Springer et al., 2025)	55.52	58.14	46.32	87.34	77.43	82.56	30.73	64.68
GRITLM (Muennighoff et al., 2024)	53.10	**61.30**	48.90	86.90	77.00	82.80	29.40	64.70
LLM2Vec (BehnamGhader et al., 2024)	55.99	58.42	45.54	87.99	76.63	84.09	29.96	64.80
Anchor(Su et al., 2025)	56.87	60.56	45.73	87.99	75.95	83.52	30.28	64.99
NV-Embed† (Lee et al., 2025a)	-	-	-	-	-	-	-	65.80
MGH (Pan et al., 2025)	57.49	58.80	47.96	87.83	**77.62**	84.04	31.10	65.87
bge-en-icl (Li et al., 2025)	**60.08**	56.67	46.55	88.51	77.31	83.69	30.68	66.08
Causal2Vec	57.28	59.46	48.89	88.43	76.41	85.38	30.57	66.10
Causal2Vec w/ ICL	57.48	59.36	**50.78**	**89.19**	77.53	**85.66**	30.82	**66.85**

Table 1: Performance comparison on MTEB (56 datasets) for models trained on publicly available retrieval datasets. `S-LLaMA-1.3B`, `Qwen2.5-1.5B`, `LLaMA-2-7B`, and `Mistral-7B` refer to embedding methods built upon these decoder-only LLMs. † denotes results reported by Pan et al. (2025). The best results are highlighted in **bold**, and the second-best are underlined. See Appendix C.8 for detailed results for each dataset.

v2 (Wang et al., 2022), a lightweight model with only 110M parameters. All LLMs are fine-tuned using LoRA (Hu et al., 2022) on A100 80GB GPUs. In particular, LoRA is also applied to the bidirectional encoder when using the 1.3B and 1.5B LLMs (see Appendix C.2 for the reason). More experimental details are provided in Appendices A and B.

4.3 MTEB Results

We evaluate the proposed Causal2Vec against competitive embedding models on the MTEB benchmark. As shown in Table 1, Causal2Vec demonstrates consistently strong performance across various LLM backbones, with our best model, Causal2Vec-Mistral-7B, achieving new state-of-the-art results compared to methods trained solely on publicly available retrieval datasets.

Comparison to Bidirectional LLM-based Methods. GRITLM (Muennighoff et al., 2024) and LLM2Vec (BehnamGhader et al., 2024) remove the causal attention mask to enable bidirectional embedding generation. Building on this attention modification, NV-Embed (Lee et al., 2025a) introduces a latent attention layer to obtain pooled embeddings, while MGH (Pan et al., 2025) initializes from LLM2Vec and leverages LLMs' inherent aggregation patterns to derive stronger embeddings over mean pooling. In contrast to these bidirectional LLM-based methods, our Causal2Vec requires no modifications to the model architecture, yet enables each token in the sequence to access contextual information through the introduced Contextual token. More importantly, our approach surpasses all aforementioned bidirectional models on MTEB. Specifically, Causal2Vec consistently outperforms LLM2Vec across three LLM base models. Notably, Causal2Vec exceeds the state-of-the-art bidirectional methods MGH (66.10 vs. 65.87) under identical training data and evaluation pipelines. These results underscore that shifting from causal attention to bidirectional is not necessary for adopting LLMs in text embedding tasks—and may even compromise the model's ability to extract the well-learned semantic information. We argue that the effectiveness of bidirectional attention in capturing contextual information relies on maintaining consistency in attention mechanisms throughout both pre-training and fine-tuning.

Comparison to Causal LLM-based Methods. Anchor (Su et al., 2025) enhances the EOS representation through two additional reconstruction objectives, but requires extensive full-parameter tuning. In comparison, our Causal2Vec outperforms Anchor by 1.11 points on both Qwen-2.5-1.5B and Mistral-7B, using only standard contrastive learning with LoRA tuning.

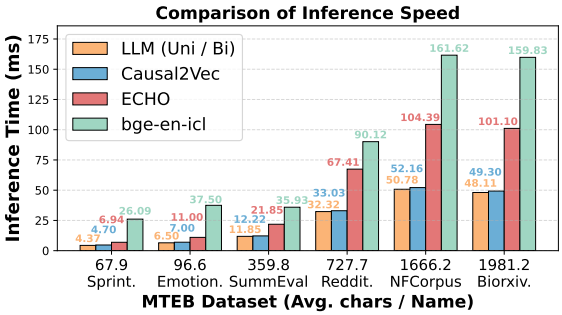


Figure 4: Average inference time per sample (in milliseconds) for various Mistral-7B-based methods on selected MTEB subsets, evaluated with batch size of 32 on a single NVIDIA A100 80GB GPU. `LLM (Uni/Bi)` denotes the standard Mistral-7B with causal or bidirectional attention. For the asymmetric dataset NFCorpus, we report the results per query-passage pair.

ECHO (Springer et al., 2025) repeats the input, allowing each token from the second occurrence to access the full sequence. However, this strategy doubles the maximum sequence length, inevitably increasing computational cost. In contrast, our Causal2Vec enables contextualized information access through a single Contextual token, and consistently outperforms ECHO on S-LLaMA-1.3B and Mistral-7B by 0.62 and 1.42 points, respectively.

The state-of-the-art unidirectional method bge-en-icl (Li et al., 2025) endows LLM-based embedding models with in-context learning (ICL) (Brown et al., 2020) capabilities by incorporating multiple task-related examples into the input. Causal2Vec achieves embedding performance on par with bge-en-icl on MTEB (66.10 vs. 66.08). However, incorporating in-context examples substantially increases the computational burden of LLMs, especially given that the maximum sequence length of bge-en-icl can reach up to 2048 tokens, which is four times that of our method. More importantly, when equipped with the same ICL strategy, Causal2Vec (w/ ICL) achieves a new state-of-the-art score of **66.85**, outperforming bge-en-icl by 0.77 points under the compute-matched setting.

Efficiency. In Figure 4, we report the approximate average inference time for different methods. Specifically, ECHO and bge-en-icl inevitably increase inference time due to their extended sequence lengths. In contrast, since the additional bidirectional encoder contains only 110M parameters and a single Contextual token adds no computational burden to the LLM, our Causal2Vec achieves inference speed comparable to the standard Mistral-7B using causal or bidirectional attention. Notably, compared to the previous best-performing method, beg-en-icl, Causal2Vec reduces inference time by up to **82%** (e.g., Sprint.: 4.37 vs. 26.09; Emotion.: 6.50 vs. 37.50). See Appendix C.7 for a comparison of the required sequence length.

CtxToken	Concat	S-LLaMA-1.3B	Mistral-7B
✗	✗	61.81	64.84
✓	✗	62.19	65.44
✓	✓	**62.83**	**65.85**

Table 2: Performance comparison of different components on MTEB-MINI (30 datasets). `CtxToken` indicates adding the Contextual token to LLM's input, while `Concat` denotes the proposed representation method that concatenates the last hidden states of LLM's Contextual and EOS tokens as the text embedding.

Method	Retr. (6)	MTEB-MINI (30)
Mistral-7B	49.44	64.84
w/ noise	44.02 (**-5.42**)	58.69 (**-6.15**)
Causal2Vec	50.81	65.85
w/ noise	47.56 (-3.25)	60.95 (-4.90)

Table 3: Performance comparison under input noise for retrieval tasks (6 datasets) and MTEB-MINI (30 datasets), between standard Mistral-7B with last-token pooling and Causal2Vec-Mistral-7B.

4.4 Ablation Studies

Effectiveness of Each Component. To evaluate the effectiveness of the proposed Contextual token and representation method, we conduct ablation studies on MTEB-MINI using two base models of different scales: S-LLaMA-1.3B and Mistral-7B. As shown in Table 2, incorporating the Contextual token into the LLM's causal attention mechanism yields average score improvements of 0.38 and 0.60 for S-LLaMA-1.3B and Mistral-7B, respectively. These results not only confirm the effectiveness of the Contextual token but also highlight its scalability and applicability across different LLMs. We attribute these performance improvements to the rich contextualized content encoded in the Contextual token, which allows preceding tokens in the sequence to access accurate sentence information even without attending to future tokens. This mitigates the inherent architectural limitations in causal attention while preserving the original autoregressive paradigm.

By concatenating the last hidden states of the Contextual and EOS tokens as the final vector rep-

Method	S-LLaMA-1.3B	Mistral-7B
Embedding Concatenation	61.92	65.02
Causal2Vec	**62.83**	**65.85**

Table 4: Performance comparison between Causal2Vec and direct embedding concatenation on MTEB-MINI (30 datasets).

Method	S-LLaMA-1.3B	Mistral-7B
Causal2Vec	**62.83**	**65.85**
w/ 2 CtxTokens	62.77	65.76
w/ 4 CtxTokens	62.51	65.72
w/ 8 CtxTokens	62.67	65.68

Table 5: Performance comparison of Causal2Vec using different numbers of Contextual tokens (`CtxTokens`) on MTEB-MINI (30 datasets). Note: Causal2Vec uses a single Contextual token by default.

resentation, we observe consistent performance improvements across all seven tasks, with average score gains of 1.02 and 1.01 on S-LLaMA-1.3B and Mistral-7B compared to standard LLMs, respectively. To further evaluate the robustness of our representation method in alleviating recency bias, we append task-irrelevant text to LLM's input sequence as noise. As shown in Table 3, standard Mistral-7B with last-token pooling suffers a significant average performance drop of 6.15 points, while Causal2Vec exhibits a smaller degradation of 4.90. Notably, the gap is larger in retrieval tasks, where the drops are 5.42 vs. 3.25, revealing the high sensitivity of last-token pooling to noise in long-text scenarios. See Appendix C.4 for further discussion on various representation methods.

To verify whether the performance gains of our method arise merely from embedding fusion, we compare Causal2Vec with a baseline that directly concatenates the output representations from the LLM and the bidirectional encoder. As shown in Table 4, this simple strategy leads to significant performance degradation, suggesting that the improvements achieved by Causal2Vec stem from addressing the inherent shortcomings of causal attention and last-token pooling.

The Number of the Contextual Tokens. To examine the impact of using multiple Contextual tokens, we adopt cross-attention with a set of learnable queries to extract a fixed number of Contextual tokens from the bidirectional encoder, following Carion et al. (2020); Li et al. (2023a). As presented in Table 5, increasing the number of the Contextual tokens leads to performance degradation. We hypothesize that the additional tokens fail to provide more distinctive semantic information and instead introduce redundancy. These findings suggest that a single Contextual token is sufficient to supply the necessary contextual information for Causal2Vec, while maintaining model simplicity and efficiency.

Method	Params.	S-LLaMA-1.3B	Mistral-7B
Causal2Vec	110M	**62.83**	**65.85**
w/ GTE-small	33M	62.71	65.45
w/ E5-small-v2	33M	62.70	65.41
w/ E5-large-v2	335M	62.31	65.71
w/o Bi-Encoder	N/A	61.81	64.84
TP (Fu et al., 2025)	N/A	61.66	64.93

Table 6: Performance comparison of Causal2Vec using different bidirectional encoders (Bi-Encoders) on MTEB-MINI (30 datasets). Note: E5-base-v2 (Wang et al., 2022) is used as the default Bi-Encoder. "Params." indicates the parameter count of the Bi-Encoder.

Impact of Different Bidirectional Encoders. Finally, we explore the impact of using different bidirectional encoders in Causal2Vec, including E5-small-v2, E5-base-v2, E5-large-v2 (Wang et al., 2022), and GTE-small (Li et al., 2023b), which achieve standalone MTEB scores of 59.93, 60.97, 61.44, and 61.36, respectively. As shown in Table 6, incorporating the Contextual token consistently outperforms the baseline LLM with last-token pooling (`w/o Bi-Encoder`), confirming the robustness of Causal2Vec across various bidirectional encoders. Notably, replacing the default E5-base-v2 with stronger encoders such as E5-large-v2 and GTE-small does not yield further meaningful performance gains. These findings suggest that the performance ceiling of our architecture is primarily determined by the LLM itself rather than the additional bidirectional encoder.

In addition, we compare Causal2Vec with TP (Fu et al., 2025), which prepends each layer's output EOS token to the input sequence of next layer. Our method substantially outperforms TP, demonstrating that compared to repeatedly using EOS token within the LLM, our Contextual token generated by the additional encoder is a more effective solution for providing contextualized information.

5 Conclusion

This paper presented Causal2Vec, a simple yet powerful text embedding model built upon decoder-only LLMs. It requires no architectural modifica-

tions or additional input text, achieving consistently strong performance in text embedding tasks. By introducing the Contextual token, we enable each token in the sequence to capture contextual information within the inherent autoregressive modeling paradigm. To address the limitations of last-token pooling, commonly used in unidirectional models, we proposed a specialized representation method that concatenates the last hidden states of the Contextual and EOS tokens as the final text embedding. Experimental results demonstrated that Causal2Vec achieves new state-of-the-art performance on MTEB across different LLM backbones. Extensive ablations and analyses further confirmed the effectiveness of the proposed mechanism.

Limitations

Despite the effectiveness of Causal2Vec, several limitations should be acknowledged: (1) Our findings suggest that a single Contextual token is sufficient to provide the missing contextual information for decoder-only LLMs. Future work could explore generating additional Contextual tokens using different bidirectional encoders or utilizing multiple task-related examples. (2) Our experiments are limited to four popular LLMs with fewer than 7B parameters, while further validation on more diverse and larger-scale LLMs could better demonstrate the scalability and robustness of our mechanism. (3) Existing state-of-the-art embedding methods (BehnamGhader et al., 2024; Springer et al., 2025; Lee et al., 2025a; Pan et al., 2025) primarily focus on evaluations using English benchmarks. We plan to further explore the performance of Causal2Vec on multilingual text embedding tasks. (4) The representation method used in Causal2Vec doubles the output embedding dimension, which may increase storage overhead in certain scenarios.

References

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

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

DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora question pairs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. 2025. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3181.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading

comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46. Association for Computational Linguistics.

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*.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Angqing Jiang, Jianlyu Chen, Zhe Fang, Yongcan Wang, Xinpeng Li, Keyu Ding, and Defu Lian. 2026. Cmedteb & care: Benchmarking and enabling efficient chinese medical retrieval via asymmetric encoders. *arXiv preprint arXiv:2604.10937*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

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

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025b. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. 2025. Making text embedders few-shot learners. In *The Thirteenth International Conference on Learning Representations*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR.

Xianming Li and Jing Li. 2024. BeLLM: Backward dependency enhanced large language model for sentence embeddings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 792–804.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. ChatQA: Surpassing GPT-4 on conversational QA and RAG. In *Advances in Neural Information Processing Systems*.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1365.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfrembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog*, 3:6.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*.

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.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2017. MS MARCO: A human-generated MAchine reading COmprehension dataset.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, volume 35, pages 27730–27744.

Tengyu Pan, Zhichao Duan, Zhenyu Li, Bowen Dong, Ning Liu, Xiuxing Li, and Jianyong Wang. 2025. Negative matters: Multi-granularity hard-negative synthesis and anchor-token-aware pooling for enhanced text embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31102–31118.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

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

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*.

Chang Su, Dengliang Shi, Siyuan Huang, Jintao Du, Changhua Meng, Yu Cheng, Weiqiang Wang, and Zhouhan Lin. 2025. Training llms to be better text embedders through bidirectional reconstruction. *arXiv preprint arXiv:2509.03020*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11897–11916.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *arXiv preprint arXiv:2507.11412*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2681–2690.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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.

Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. 2026a. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. In *Proceedings of the ACM Web Conference 2026*, pages 1971–1982.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026b. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023a. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Experimental Details for Training

A.1 Implementation Details

In this section, we provide additional experimental details based on Section 4.2. Following the implementation of open-source LLM2Vec, we employ the same fixed random seed to guarantee fair comparison and reproducibility of our results. In addition, we use the AdamW optimizer with an initial learning rate of 1e-4, and a warm-up strategy for the first 300 steps followed by linear decay over the remaining steps. To reduce GPU memory usage, all models are trained with bfloat16 quantization, gradient checkpointing, and FlashAttention-2 (Dao, 2024). We also apply gradient accumulation to process a large batch size of 512 while sampling the same dataset within each batch. The maximum sequence length is set to the standard 512 tokens by default.

Hyperparameter	1.3 B & 1.5 B	7B
Batch Size	128	64
Gradient Accumulation Steps	4	8
Training Steps	2000	1000
Maximum Steps	4000	2000
LoRA Rank	64	64
LoRA Alpha	128	32
LoRA for Bidirectional Encoder	✓	✗

Table 7: Hyperparameters used in the experiments.

Table 7 presents the hyperparameters that vary across different base models. We set the LoRA (Hu

Dataset	Instruction (s)
ELI5	Provided a user question, retrieve the highest voted answers on Reddit ELI5 forum
HotpotQA	Given a multi-hop question, retrieve documents that can help answer the question
FEVER	Given a claim, retrieve documents that support or refute the claim
MIRACL	Given a question, retrieve Wikipedia passages that answer the question
MSMARCO Passage	Given a web search query, retrieve relevant passages that answer the query
MSMARCO Document	Given a web search query, retrieve relevant documents that answer the query
NQ	Given a question, retrieve Wikipedia passages that answer the question
NLI	Given a premise, retrieve a hypothesis that is entailed by the premise Retrieve semantically similar text
SQuAD	Retrieve Wikipedia passages that answer the question
TriviaQA	Retrieve Wikipedia passages that answer the question
QuoraDuplicates	Given a question, retrieve questions that are semantically equivalent to the given question Find questions that have the same meaning as the input question
Mr. TyDi	Given a question, retrieve Wikipedia passages that answer the question
DuReader	Given a Chinese search query, retrieve web passages that answer the question
T2Ranking	Given a Chinese search query, retrieve web passages that answer the question

Table 8: Instructions used for public retrieval datasets.

Category	Dataset
Retrieval (6)	SciFact, NFCorpus, FiQA2018, SCIDOCS, TRECCOVID, Touche2020
Reranking (3)	AskUbuntuDupQuestions, SciDocsRR, StackOverflowDupQuestions
Clustering (5)	BiorxivClusteringS2S, BiorxivClusteringP2P, MedrxivClusteringS2S, MedrxivClusteringP2P, TwentyNewsgroupsClustering
Pair Classification (2)	SprintDuplicateQuestions, TwitterURLCorpus
Classification (6)	AmazonReviewsClassification, Banking77Classification, EmotionClassification, MTOPDomainClassification, TweetSentimentExtractionClassification, ImdbClassification
STS (7)	BIOSSES, STS12, STS13, STS14, STS15, STS16, STSBenchmark
SummEval (1)	SummEval
Overall	30 datasets

Table 9: Composition of the MTEB-MINI benchmark.

et al., 2022) rank to 64 and LoRA alpha to 32 for 7B models following Li et al. (2025), while a large LoRA alpha of 128 is used for smaller models. For training steps, 7B models are trained for 1000 steps as in BehnamGhader et al. (2024); Pan et al. (2025), whereas smaller models require 2000 steps. The maximum steps for the linear scheduler are set to twice the training steps. In terms of computational cost, training a 7B model typically requires about 33 A100 80GB GPU hours, while the 1.3B and 1.5B models require over 20 GPU hours.

As we adopt instruction-tuned versions of Qwen2.5-1.5B, LLaMA-2-7B, and Mistral-7B, special tokens are added to the input text according to the official instruction prompt templates for each model. Specifically, `[INST]` and `[/INST]` are used for LLaMA-2-7B and Mistral-7B, while `<|im_start|>user\n` and `<|im_end|>` are added for Qwen2.5-1.5B. Moreover, to insert the Contextual token into LLM's input sequence, we first add a placeholder to the input text: `<s>` for S-LLaMA-1.3B, LLaMA-2-7B, and Mistral-7B, and `<|end_of_text|>` for Qwen2.5-1.5B. After tokenizing the modified input and obtaining its word embeddings from LLM's embedding layer, we replace the placeholder with the Contextual token.

A.2 HuggingFace Models

All base models and bidirectional encoders employed in this work are obtained from the HuggingFace platform:

Base Models

- S-LLaMA-1.3B (Xia et al., 2024): princeton-nlp/Sheared-LLaMA-1.3B
- Qwen2.5-1.5B (Team, 2024a,b): Qwen/Qwen2.5-1.5B-Instruct
- LLaMA-2-7B (Touvron et al., 2023): meta-llama/Llama-2-7b-chat-hf
- Mistral-7B (Jiang et al., 2023): mistralai/Mistral-7B-Instruct-v0.2

Bidirectional Encoders

- GTE-small (Li et al., 2023b): thenlper/gte-small
- E5-small-v2 (Wang et al., 2022): intfloat/e5-small-v2
- E5-base-v2 (Wang et al., 2022): intfloat/e5-base-v2
- E5-large-v2 (Wang et al., 2022): intfloat/e5-large-v2

A.3 Public Retrieval Datasets and Instructions

The collection of publicly available retrieval datasets used for training is curated by Springer et al. (2025) and is distributed under the Apache License 2.0. It includes the following datasets: ELI5 (sample ratio 0.1) (Fan et al., 2019), HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), MIRACL (Zhang et al., 2023b), MS-MARCO passage ranking (sample ratio 0.5) and document ranking (sample ratio 0.2) (Nguyen et al., 2017), NQ (Karpukhin et al., 2020), NLI (Gao et al., 2021), SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), Quora Duplicate Questions (sample ratio 0.1) (DataCanary et al., 2017), Mr. TyDi (Zhang et al., 2021), DuReader (He et al., 2018), and T2Ranking (sample ratio 0.5) (Xie et al., 2023).

Table 8 lists the instructions used for each dataset, which are manually written by Wang et al. (2024a). Notably, in the query-passage pairs of retrieval datasets, task-specific instructions are appended only to the queries, without modifying the passages.

B Experimental Details for Evaluation

B.1 MTEB-MINI Details

Considering the substantial computational resources required for full evaluation on MTEB, we follow Springer et al. (2025); Pan et al. (2025) and select a subset of MTEB for ablation and analysis. While prior studies (BehnamGhader et al., 2024; Su et al., 2025) utilize only a few datasets, our preliminary experiments suggest that evaluation on a limited subset may introduce significant bias and fail to effectively reflect the overall trends of the full MTEB. To this end, as shown in Table 9, we empirically introduce the MTEB-MINI by selecting 30 representative datasets spanning all task categories in MTEB.

B.2 Evaluation Metrics

The task categories of MTEB include Retrieval (Retr.), Reranking (Rerank.), Clustering (Clust.), Pair Classification (PairClass.), Classification (Class.), Semantic Textual Similarity (STS), and Summarization (Summ.). For these tasks, the main evaluation metrics are nDCG@10, MAP, V-measure (V-meas.), average precision (AP), accuracy (Acc.), and Spearman correlation (Spear., both for STS and Summ.), respectively.

B.3 Instructions for MTEB Evaluation

To enable a fair comparison with prior leading embedding methods (Wang et al., 2024a; Springer et al., 2025; BehnamGhader et al., 2024; Lee et al., 2025a; Li et al., 2025; Pan et al., 2025), we use the same instruction prompts for evaluation on both MTEB and MTEB-MINI. The instructions applied to each dataset are listed in Table 16.

C Additional Results

C.1 The L2 Norms of Contextual and EOS Tokens

To examine the respective contributions of Contextual and EOS tokens to the final text embedding, we compare their L2 norms on selected MTEB datasets. As depicted in Figure 5, we observe that the EOS token consistently shows higher L2 norms across various task categories, indicating its greater influence on the concatenated representation.

Method	S-LLaMA-1.3B	Qwen2.5-1.5B	Mistral-7B
w/ Bi-LoRA	**62.83**	**64.05**	65.64
w/o Bi-LoRA	62.57	63.82	**65.85**

Table 10: Performance comparison on MTEB-MINI (30 datasets) between Causal2Vec with and without applying LoRA to the bidirectional encoder (`Bi-LoRA`), using three base models: S-LLaMA-1.3B, Qwen2.5-1.5B, and Mistral-7B.

C.2 Impact of Freezing the Bidirectional Encoder

Since the BERT-style bidirectional encoder (E5-base-v2) we use is specifically trained for embedding tasks, this section investigates whether it should be frozen during fine-tuning. As shown in Table 10, we observe that fine-tuning the bidirectional encoder with LoRA leads to an average score improvement of 0.26 and 0.23 on the MTEB-MINI for S-LLaMA-1.3B and Qwen2.5-1.5B, respectively, but degrades Mistral-7B's performance

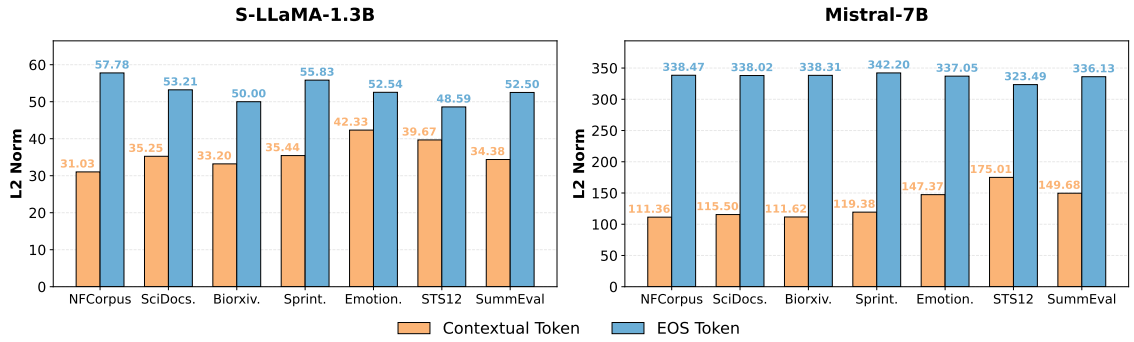


Figure 5: L2 norms of Contextual and EOS tokens on selected MTEB subsets for two base models: S-LLaMA-1.3B and Mistral-7B. The evaluated datasets span seven tasks, including NFCorpus, SciDocsRR (SciDocs.), BiorxivClusteringP2P (Biorxiv.), SprintDuplicateQuestions (Sprint.), EmotionClassification (Emotion.), STS12, and SummEval.

Method	Retr.	MTEB-MINI
e5-mistral-7b-instruct	49.10	65.75
Causal2Vec	51.22 (+2.12)	66.59 (+0.84)

Table 11: Average scores on the retrieval task (6 datasets) and MTEB-MINI (30 datasets) for e5-mistral-7b-instruct and Causal2Vec$_{\text{e5-mistral-7b-instruct}}$.

by 0.21 points. We attribute this to two potential effects of making the bidirectional encoder trainable: (1) it may cause catastrophic forgetting in the bidirectional encoder, and (2) it may help the LLM better interpret the Contextual token through joint fine-tuning. Large-scale LLMs are more susceptible to the former, as they already have sufficient capacity to comprehend newly added tokens during fine-tuning (Liu et al., 2023; Li et al., 2023a). This suggests that whether the introduced bidirectional encoder should remain frozen may depend on the scale of the underlying LLM.

C.3 Effectiveness on Existing Embedding Models

In this section, we examine the effectiveness of Causal2Vec on existing high-performing embedding models rather than on general-purpose LLMs. As shown in Table 11, Causal2Vec remains effective when applied to e5-mistral-7b-instruct (Wang et al., 2024a), a powerful embedding model that uses causal attention with last-token pooling and is fine-tuned on extensive synthetic data. This demonstrates that our method indeed mitigates the inherent limitations of causal attention in representation learning. Moreover, we observe larger improvements on retrieval tasks involving long text inputs, further confirming the effectiveness of our proposed representation method.

C.4 Different Representation Methods

In this section, we explore various representation methods tailored to our embedding framework through the following experimental settings (Note: all experiments incorporate the Contextual token into the LLM's input sequence):

1. **Default:** Causal2Vec generates the text embedding by concatenating LLM's output hidden states of the Contextual and EOS tokens.
2. **Concat-bi:** Concatenate the LLM's output EOS token with the output of the bidirectional encoder processed by an MLP layer.
3. **Average:** Take the average of the LLM's output hidden states corresponding to the Contextual and EOS tokens.
4. **Last-token pooling:** Use only the final hidden states of the last EOS token as text embedding.

Method	S-LLaMA-1.3B	Mistral-7B
Causal2Vec	**62.83**	**65.85**
w/ Concat-bi	62.44	65.59
w/ Average	62.67	65.55
w/ Last-token	62.19	65.44

Table 12: Performance comparison on MTEB-MINI (30 datasets) using different representation method with two base models: S-LLaMA-1.3B and Mistral-7B.

Table 12 presents the comparison results on MTEB-MINI, from which we draw the following observations: (1) The representation methods that

Method	Dim.	ArguAna	SciFace	NFCorpus
Default	8192	324.08s	243.10s	173.81s
Average	4096	322.70s	242.02s	173.27s

Table 13: End-to-end completion time (in seconds) for $\text{Causal2Vec}_{\text{mistral-7b}}$ across retrieval datasets with different output embedding dimensions.

Method	S-LLaMA-1.3B	Mistral-7B
before instruction	62.69	65.55
after instruction	**62.83**	**65.85**

Table 14: Average MTEB-MINI (30 datasets) score for placing the Contextual token before and after the task-specific instruction in Causal2Vec.

utilize both the EOS and Contextual tokens consistently outperform last-token pooling for S-LLaMA-1.3B and Mistral-7B. This suggests that incorporating an additional context-aware token with the EOS token leads to richer semantic information while reducing the model's reliance on the single EOS token alone. (2) We observe further performance improvements when the concatenated Contextual token is derived from the LLM, rather than from the bidirectional encoder followed by an MLP layer. We speculate that this helps the LLM better capture the semantic content encoded in the Contextual token. (3) Among the strategies utilizing the last hidden states of both EOS and Contextual tokens, concatenation yields substantially better text representations than averaging.

We further compare the overall completion time across retrieval datasets using the concatenation and averaging representation methods. As shown in Table 13, although Causal2Vec doubles the output embedding dimension under concatenation, the additional runtime overhead is negligible, suggesting that LLM inference remains the dominant computational bottleneck in practice.

C.5 The Position of the Contextual Token.

We investigate whether the position of the Contextual token affects embedding performance by comparing two placement settings: before vs. after the instruction. As shown in Table 14, "`after instruction`" consistently yields better results. We speculate that positioning the Contextual token before the instruction tokens may hinder the LLM's ability to accurately interpret and follow task-specific prompts.

Attention	S-LLaMA-1.3B	LLaMA-2-7B	Mistral-7B
Causal	**62.83**	**64.62**	**65.85**
Bidirectional	62.52	64.30	65.77

Table 15: Performance comparison of Causal2Vec using bidirectional and causal attention mechanisms on MTEB-MINI (30 datasets), with three base models: S-LLaMA-1.3B, LLaMA-2-7B, and Mistral-7B.

C.6 Impact of Different Attention Mechanisms

This section verifies the impact of different attention mechanisms in our proposed Causal2Vec. As illustrated in Table 15, shifting from causal to bidirectional attention consistently leads to lower performance of on MTEB-MINI, even for the Mistral-7B backbone, which generally benefits from attention modifications in most bidirectional attention-based embedding methods (Muennighoff et al., 2024; BehnamGhader et al., 2024; Lee et al., 2025a; Pan et al., 2025). We hypothesize that the attention mismatch between pre-training and contrastive learning compromises LLM's ability to extract the well-learned semantic information. This finding further highlights that altering the original attention mechanism of LLMs may be suboptimal for text embedding tasks.

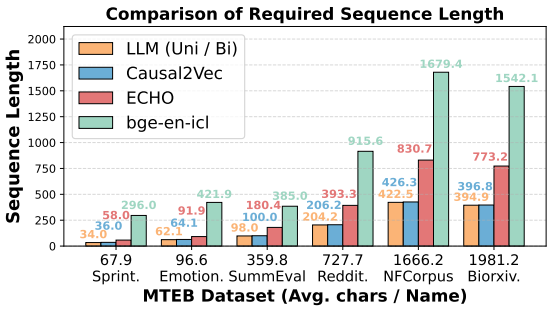


Figure 6: Average required sequence length per sample for various Mistral-7B-based methods on selected MTEB subsets. `LLM (Uni/Bi)` denotes the standard Mistral-7B with causal or bidirectional attention. For the asymmetric dataset NFCorpus, we report the results per query-passage pair. Note: Echo (Springer et al., 2025) repeats only the input text, excluding the instruction. For bge-en-icl (Li et al., 2025), in-context examples are taken from the official repository.

C.7 Comparison on Required Sequence Length

Figure 6 presents the approximate average sequence lengths required by different models on selected MTEB subsets. Specifically,

ECHO (Springer et al., 2025) repeats the input, while bge-en-icl (Li et al., 2025) incorporates several task-related in-context examples, both of which significantly increase the required sequence length for LLMs. In contrast, our method requires only a single additional Contextual token compared to standard LLMs. Notably, when using <s> as a placeholder for the Contextual token in Mistral-7B, tokenization produces not only the <s> token itself but also an additional separator token. As a result, Causal2Vec reduces the required sequence length by up to **85%** (Sprint.: 34.0 vs. 269.0; Emotion.: 62.1 vs. 421.9) compared to the state-of-the-art unidirectional method bge-en-icl.

C.8 Full MTEB Results

We present detailed results on all 56 MTEB datasets for the proposed Causal2Vec in Table 17 and Table 18, including four base models: S-LLaMA-1.3B, Qwen2.5-1.5B, LLaMA-2-7B, and Mistral-7B.

D Related Works

Many industry-developed embedding models, such as KaLM-Embedding (Hu et al., 2025), Gemini Embedding (Lee et al., 2025b), and Qwen3 Embedding (Zhang et al., 2025), have achieved remarkably outstanding performance on MTEB. However, these approaches heavily rely on extensive proprietary synthetic data for training as well as various engineering optimizations, making direct comparison with academic research unfair (Li et al., 2025; Su et al., 2025). Therefore, we do not include these models in our comparisons, we instead briefly introduce them to highlight their contributions.

KaLM-Embedding (Hu et al., 2025) is a superior embedding model that aims to improve the training data quality and distills knowledge from LLMs into text embeddings. Specifically, it is trained on a carefully curated corpus that includes more than 20 data categories for pre-training and 70 categories for fine-tuning. In addition, KaLM-Embedding applies several key techniques (Dai et al., 2022; Meng et al., 2024; Wang et al., 2024a) to further enhance and clean the data, resulting in a substantially larger and higher-quality training dataset.

Gemini Embedding (Lee et al., 2025b) is built upon the powerful Gemini LLM (Team et al., 2024) and trained on a wide range of embedding tasks. It leverages Gemini to guide the construction of a diverse and high-quality training dataset. Furthermore, the final embedding model is obtained by combining several fine-tuned checkpoints using an effective parameter-averaging technique (Wortsman et al., 2022).

Qwen3 Embedding series (Zhang et al., 2025) are built upon the Qwen3 foundation models (Yang et al., 2025) and exhibit strong performance in text embedding and reranking. The training pipeline of Qwen3 Embedding combines large-scale unsupervised pre-training with supervised fine-tuning on extensive training datasets. In particular, the high-quality and diverse training data covering multiple domains and languages is synthesized by the Qwen3 LLMs.

E Ethical Considerations

Our proposed Causal2Vec can be applied to a wide range of real-world applications, including information retrieval and LLM-based retrieval-augmented generation systems. However, LLMs are known to suffer from biases and hallucinations, which could potentially lead to negative social impacts.

Dataset	Instruction Template
AmazonCounterfactualClassification	Classify a given Amazon customer review text as either counterfactual or not-counterfactual.
AmazonPolarityClassification	Classify Amazon reviews into positive or negative sentiment
AmazonReviewsClassification	Classify the given Amazon review into its appropriate rating category
Banking77Classification	Given a online banking query, find the corresponding intents
EmotionClassification	Classify the emotion expressed in the given Twitter message into one of the six emotions: anger, fear, joy, love, sadness, and surprise.
ImdbClassification	Classify the sentiment expressed in the given movie review text from the IMDB dataset.
MassiveIntentClassification	Given a user utterance as query, find the user intents
MassiveScenarioClassification	Given a user utterance as query, find the user scenarios
MTOPDomainClassification	Classify the intent domain of the given utterance in task-oriented conversation
MTOPIntentClassification	Classify the intent of the given utterance in task-oriented conversation
ToxicConversationsClassif.	Classify the given comments as either toxic or not toxic
TweetSentimentClassification	Classify the sentiment of a given tweet as either positive, negative, or neutral
ArxivClusteringP2P	Identify the main and secondary category of Arxiv papers based on the titles and abstracts.
ArxivClusteringS2S	Identify the main and secondary category of Arxiv papers based on the titles
BiorxivClusteringP2P	Identify the main category of Biorxiv papers based on the titles and abstracts
BiorxivClusteringS2S	Identify the main category of Biorxiv papers based on the titles
MedrxivClusteringP2P	Identify the main category of Medrxiv papers based on the titles and abstracts
MedrxivClusteringS2S	Identify the main category of Medrxiv papers based on the titles
RedditClustering	Identify the topic or theme of Reddit posts based on the titles
RedditClusteringP2P	Identify the topic or theme of Reddit posts based on the titles and posts
StackExchangeClustering	Identify the topic or theme of StackExchange posts based on the titles
StackExchangeClusteringP2P	Identify the topic or theme of StackExchange posts based on the given paragraphs
TwentyNewsgroupsClustering	Identify the topic or theme of the given news articles
SprintDuplicateQuestions	Retrieve duplicate questions from Sprint forum
TwitterSemEval2015	Retrieve tweets that are semantically similar to the given tweet
TwitterURLCorpus	Retrieve tweets that are semantically similar to the given tweet
AskUbuntuDupQuestions	Retrieve duplicate questions from AskUbuntu forum
MindSmallReranking	Retrieve relevant news articles based on user browsing history
SciDocsRR	Given a title of a scientific paper, retrieve the titles of other relevant papers
StackOverflowDupQuestions	Retrieve duplicate questions from StackOverflow forum
ArguAna	Given a claim, find documents that refute the claim
ClimateFEVER	Given a claim about climate change, retrieve documents that support or refute the claim.
CQADupstackRetrieval	Given a question, retrieve detailed question descriptions from Stackexchange that are duplicates to the given question.
DBPedia	Given a query, retrieve relevant entity descriptions from DBPedia
FEVER	Given a claim, retrieve documents that support or refute the claim
FiQA2018	Given a financial question, retrieve user replies that best answer the question
HotpotQA	Given a multi-hop question, retrieve documents that can help answer the question
MSMARCO	Given a web search query, retrieve relevant passages that answer the query
NFCorpus	Given a question, retrieve relevant documents that best answer the question
NQ	Given a question, retrieve Wikipedia passages that answer the question
QuoraRetrieval	Given a question, retrieve questions that are semantically equivalent to the given question.
SCIDOCS	Given a scientific paper title, retrieve paper abstracts that are cited by the given paper
SciFact	Given a scientific claim, retrieve documents that support or refute the claim
Touche2020	Given a question, retrieve detailed and persuasive arguments that answer the question
TRECCOVID	Given a query on COVID-19, retrieve documents that answer the query
STS*	Retrieve semantically similar text.
SummEval	Given a news summary, retrieve other semantically similar summaries

Table 16: Instructions used for evaluation on the MTEB. "STS*" denotes that the corresponding instruction is applied to all STS datasets.

Dataset	S-LLaMA-1.3B	Qwen2.5-1.5B	LLaMA-2-7B
AmazonCounterfactualClassification	74.49	73.04	76.79
AmazonPolarityClassification	92.75	94.23	94.80
AmazonReviewsClassification	46.48	46.61	51.75
ArguAna	54.73	58.59	57.35
ArxivClusteringP2P	46.25	49.16	48.37
ArxivClusteringS2S	39.52	44.97	42.88
AskUbuntuDupQuestions	61.59	64.18	63.54
BIOSSES	83.96	85.94	84.06
Banking77Classification	85.96	86.74	88.14
BiorxivClusteringP2P	38.13	38.71	39.05
BiorxivClusteringS2S	35.13	36.80	36.42
CQADupstackRetrieval	39.53	43.77	43.42
ClimateFEVER	32.62	34.70	32.46
DBPedia	44.85	45.88	49.85
EmotionClassification	46.82	48.52	49.74
FEVER	88.11	90.02	90.53
FiQA2018	44.52	48.41	51.29
HotpotQA	67.13	68.43	71.45
ImdbClassification	83.76	83.71	88.33
MSMARCO	40.88	41.41	41.22
MTOPDomainClassification	94.05	94.60	95.53
MTOPIntentClassification	73.45	78.37	82.38
MassiveIntentClassification	73.36	75.78	77.63
MassiveScenarioClassification	77.58	78.91	79.88
MedrxivClusteringP2P	33.38	34.31	33.13
MedrxivClusteringS2S	31.58	32.64	32.14
MindSmallReranking	32.71	32.47	32.46
NFCorpus	37.43	39.59	40.21
NQ	58.02	59.51	64.10
QuoraRetrieval	88.91	89.39	88.80
RedditClustering	56.91	57.66	63.07
RedditClusteringP2P	61.26	60.78	64.31
SCIDOCS	19.56	20.64	21.28
SICK-R	81.99	82.70	82.78
STS12	77.04	79.63	78.77
STS13	87.47	88.46	88.89
STS14	83.21	84.27	85.29
STS15	88.93	89.64	89.86
STS16	86.83	87.43	87.72
STS17	91.13	91.53	92.19
STS22	69.21	68.16	70.67
STSBenchmark	87.85	88.53	88.77
SciDocsRR	81.68	83.74	84.11
SciFact	73.04	74.67	75.77
SprintDuplicateQuestions	96.26	96.57	97.00
StackExchangeClustering	64.27	67.86	69.14
StackExchangeClusteringP2P	32.41	35.09	36.74
StackOverflowDupQuestions	50.16	52.70	52.61
SummEval	31.45	30.99	31.09
TRECCOVID	76.24	76.97	78.65
Touche2020	24.74	26.49	22.83
ToxicConversationsClassification	65.03	62.07	65.02
TweetSentimentExtractionClassification	61.53	61.61	61.44
TwentyNewsgroupsClustering	49.01	51.29	54.33
TwitterSemEval2015	75.24	76.54	79.78
TwitterURLCorpus	87.03	87.05	86.77
MTEB Average (56)	62.63	63.97	64.94

Table 17: Results of Causal2Vec on all 56 MTEB datasets across three base models: S-LLaMA-1.3B, Qwen2.5-1.5B, and LLaMA-2-7B.

Dataset	Mistral-7B	Mistral-7B (w/ ICL)
AmazonCounterfactualClassification	76.22	75.99
AmazonPolarityClassification	95.02	95.80
AmazonReviewsClassification	51.40	53.78
ArguAna	57.55	59.11
ArxivClusteringP2P	48.99	50.64
ArxivClusteringS2S	45.51	47.09
AskUbuntuDupQuestions	65.96	65.71
BIOSSES	86.42	87.28
Banking77Classification	88.62	88.85
BiorxivClusteringP2P	39.24	40.82
BiorxivClusteringS2S	38.32	39.09
CQADupstackRetrieval	45.59	46.82
ClimateFEVER	35.55	34.12
DBPedia	51.65	52.09
EmotionClassification	50.56	51.40
FEVER	91.53	91.68
FiQA2018	54.96	56.03
HotpotQA	74.25	73.11
ImdbClassification	91.24	92.45
MSMARCO	42.22	42.83
MTOPDomainClassification	95.79	96.36
MTOPIntentClassification	83.12	86.23
MassiveIntentClassification	78.14	79.53
MassiveScenarioClassification	81.27	82.40
MedrxivClusteringP2P	34.33	36.32
MedrxivClusteringS2S	34.28	34.73
MindSmallReranking	32.32	32.31
NFCorpus	41.63	41.41
NQ	66.65	66.50
QuoraRetrieval	89.35	89.24
RedditClustering	64.73	64.99
RedditClusteringP2P	66.43	68.21
SCIDOCS	22.40	22.76
SICK-R	83.49	83.33
STS12	79.37	79.71
STS13	88.69	89.75
STS14	85.43	85.80
STS15	90.76	90.92
STS16	88.26	88.38
STS17	92.47	92.31
STS22	69.44	69.20
STSBenchmark	89.47	89.96
SciDocsRR	84.40	84.61
SciFact	77.52	77.92
SprintDuplicateQuestions	96.70	97.16
StackExchangeClustering	72.23	76.40
StackExchangeClusteringP2P	37.73	40.63
StackOverflowDupQuestions	55.16	54.81
SummEval	30.57	30.82
TRECCOVID	83.48	83.09
Touche2020	24.86	25.51
ToxicConversationsClassification	63.05	65.14
TweetSentimentExtractionClassification	62.52	62.46
TwentyNewsgroupsClustering	56.01	59.62
TwitterSemEval2015	81.35	82.91
TwitterURLCorpus	87.23	87.50
MTEB Average (56)	66.10	66.85

Table 18: Results of Causal2Vec on all 56 MTEB datasets for Mistral-7B and Mistral-7B (w/ ICL).