CoT-BERT: Enhancing Unsupervised Sentence Representation through Chain-of-Thought

Bowen Zhang, Kehua Chang, and Chunping Li

School of Software, Tsinghua University
zbw23@mails.tsinghua.edu.cn
changkehua@gmail.com
cli@tsinghua.edu.cn

Abstract. Unsupervised sentence representation learning aims to transform input sentences into fixed-length vectors enriched with intricate semantic information while obviating the reliance on labeled data. Recent strides within this domain have been significantly propelled by breakthroughs in contrastive learning and prompt engineering. Despite these advancements, the field has reached a plateau, leading some researchers to incorporate external components to enhance the quality of sentence embeddings. Such integration, though beneficial, complicates solutions and inflates demands for computational resources. In response to these challenges, this paper presents CoT-BERT, an innovative method that harnesses the progressive thinking of Chain-of-Thought reasoning to tap into the latent potential of pre-trained models like BERT. Additionally, we develop an advanced contrastive learning loss function and propose a novel template denoising strategy. Rigorous experimentation demonstrates that CoT-BERT surpasses a range of well-established baselines by relying exclusively on the intrinsic strengths of pre-trained models. [1]

Keywords: Sentence Representation · Self-Supervised Learning · Contrastive Learning · Chain-of-Thought.

1 Introduction

Sentence representation is a fundamental aspect of Natural Language Processing (NLP), which concentrates on converting input sentences into fixed-length numerical vectors, commonly known as sentence embeddings. These vectors are crucial for enabling computational systems and advanced deep learning models to process linguistic data effectively. Their applications span across various domains, including information retrieval, text clustering, topic modeling, recommendation systems, and the development of artificial intelligence agents [1]. Moreover, they serve as essential features for neural networks engaged in downstream tasks such as relationship extraction, sentiment classification, and text implication recognition.

---

[1] Our code and checkpoints are available at https://github.com/ZBWpro/CoT-BERT.

In this research field, unsupervised sentence representation learning has garnered substantial attention owing to its independence from annotated data. Currently, the prevailing paradigm in academia involves leveraging pre-trained language models (PLMs) like BERT [2] and RoBERTa [3], with contrastive learning to mitigate the issue of semantic space anisotropy [4] by drawing similar samples closer while pushing dissimilar ones apart.

With the rise of prompt engineering, a series of studies have been initiated to integrate prompts into sentence representation tasks. A remarkable endeavor, PromptBERT [5], adopts a manual template "This sentence: '[X]' means [MASK]." to encapsulate the input sentence [X], using the output vector corresponding to the [MASK] token as the original sentence's representation. Despite its impressive performance, subsequent advancements have encountered a plateau. For instance, ConPVP [6], an improvement on PromptBERT, achieves a marginal gain, with only a 0.03 increment in the average Spearman's correlation score across seven Semantic Textual Similarity (STS) benchmarks when implemented with RoBERTa$_{base}$.

In pursuit of further enhancements, some research [7–10] has explored incorporating external models or corpora to assist in sentence embedding derivation. Although these additions can improve model performance, they concurrently increase solution complexity and computational resource consumption. For example, the current state-of-the-art (SOTA) work, RankCSE [8], necessitates knowledge distillation from two high-capacity teacher models for training. Specifically, the development of RankCSE-BERT$_{base}$ involves training both SimCSE-BERT$_{base}$ [11] and SimCSE-BERT$_{large}$, constraining its applicability due to the computational demands of updating BERT$_{large}$. Another cutting-edge strategy, RankEncoder [9], requires the introduction of a large external corpus on top of the training set, which poses challenges in data-scarce scenarios.

This paper endeavors to fully unleash the potential of pre-trained models, achieving results on par with or superior to previous methods without introducing any external components. To this end, we draw inspiration from the progressive thinking of Chain-of-Thought (CoT) [12], attempting to derive sentence representations through a multi-staged process.

CoT, an emerging prompting technique, has shown promise in augmenting the performance of high-parameter generative models in complex reasoning tasks by breaking down problems into a series of logical steps leading to the final answer. Although BERT operates with significantly fewer parameters compared to models such as LLaMA [13] and GPT [14], we posit that the progressive nature of CoT can be adapted to discriminative models. Given BERT's capability to assign context-specific meanings to the special token [MASK] via its attention mechanism during the masked language modeling (MLM) task, this token exhibits an adaptive characteristic that aligns well with the CoT methodology.

Accordingly, we propose a two-step approach to sentence representation, which includes both comprehension and summarization stages, with the latter's output serving as the representation for the original sentence. Extensive experimental evaluations indicate that this method, termed CoT-BERT, out-

performs several robust baselines without necessitating additional parameters. The primary contributions of this paper are outlined as follows:

– **Introduction of CoT-BERT**: This study unveils CoT-BERT, a groundbreaking method that integrates the progressive logic of Chain-of-Thought with sentence representation. Our approach demonstrates performance that matches or exceeds current rank-based SOTA solutions, while obviating the need for external corpora or auxiliary text representation models. Notably, employing RoBERTa$_{base}$ as the PLM, we attain a Spearman correlation of 80.62% across seven STS tasks, markedly surpassing the existing best result. To our knowledge, CoT-BERT represents the inaugural effort to amalgamate CoT reasoning with sentence representation.
– **Extended InfoNCE Loss**: We present a superior contrastive learning loss function that extends beyond the conventional InfoNCE Loss [15]. This function facilitates an enhanced optimization of the PLM's semantic space uniformity by introducing contrast not only between anchor sentences and negative samples but also between positive and negative instances.
– **Innovative Template Denoising Strategy**: Our research proposes an advanced template denoising strategy to eliminate the potential influence of semantic interpretation attributable to prompt biases. This objective is realized by filling blank templates with [PAD] placeholders of identical length to the input sentence and adjusting the attention masks accordingly.
– **Comprehensive Experimental Evaluation**: We conduct a thorough examination of CoT-BERT's performance across seven established STS benchmarks. Additionally, we provide exhaustive ablation studies and analyses focused on the main innovations of CoT-BERT to ascertain the reasons behind its effectiveness. Furthermore, our code and checkpoints have been made available for replication and further experimentation.

2   Related Work

This section reviews three categories of research directly related to our work: unsupervised contrastive learning methods for sentence embeddings, techniques for text representation through prompts, and chain-of-thought prompting for multi-step reasoning.

2.1   Unsupervised Sentence Representation

Inspired by advancements in computer vision, numerous studies have explored the application of contrastive learning to unsupervised sentence representation [17–20]. Compared to post-processing strategies like BERT-flow [4] and BERT-whitening [21], contrastive learning demonstrates more pronounced improvements within the semantic space of BERT, thus establishing itself as the predominant method for deriving embeddings in the current NLP community. Among these efforts, ConSERT [16] employs four strategies, including adversarial attacks and token shuffling, to construct positive samples. SimCSE [11], on the

other hand, utilizes standard dropout for minimal data augmentation. Building upon this, SSCL [22] leverages outputs from the intermediate layers of PLMs as hard negatives to mitigate over-smoothing. In general, these methods primarily innovate in the construction of positive and negative samples, yet they do not fully exploit BERT's behavior during its pre-training phase.

A new frontier in this field involves integrating external components to refine sentence embeddings. DiffCSE [7], for instance, introduces a generator and discriminator structure atop the encoder, aiming to endow the model with the ability to distinguish between original and edited sentences. RankCSE [8] proposes utilizing multiple text representation models for knowledge distillation, coupled with a training methodology that combines ranking consistency and contrastive learning. Moreover, RankEncoder [9] suggests the inclusion of an external corpus to enrich sentence representation derivation. While methods that incorporate external models or datasets typically exhibit enhanced performance over those relying solely on PLMs, this integration inevitably increases the complexity and resource demands of their frameworks.

2.2   Prompt-based Learning

Originating from GPT models [14], the concept of prompts has rapidly expanded into diverse domains, including semantic textual similarity. Prompt-based techniques are designed to maximize the utilization of prior knowledge stored within PLMs. For models like BERT and RoBERTa, this objective is chiefly realized through the transformation of downstream tasks into formats that closely emulate MLM. Notably, PromptBERT [5] employs a manually designed template to encapsulate the input sentence and takes the output vector corresponding to the [MASK] token as the representation. In a parallel endeavor, PromCSE [23] deploys a distinctive approach by freezing the entire pre-trained model while integrating multi-layer learnable soft prompts. Furthermore, ConPVP [6] merges continuous prompts with various manual templates. These elements are then converted into anchor sentences, positive instances, and negative instances for incorporation into the InfoNCE Loss.

2.3   Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting represents a revolutionary approach, guiding large language models (LLMs) through a series of intermediate steps towards the final answer. As CoT is primarily targeted at direct inference scenarios and typically exhibits significant advantages only when the model reaches a substantial size, CoT reasoning is considered an emergent property of LLMs [12].

Despite this, the principle of decomposing complex problem, as advocated by CoT, finds broad applicability in deep learning. As an illustration, consider the representation of text in neural networks, wherein the journey commences with the encoding of individual lexical units before advancing toward holistic sentence representation. Therefore, we aspire to extend this technique to discriminative models like BERT to further unlock the potential inherent within PLMs.

3   Methodology

In this section, we initially present our CoT-style two-stage sentence representation strategy, along with its underlying design principles, in subsection 3.1. Following this, in subsection 3.2, we proceed to introduce the extended InfoNCE Loss, accompanied by an intuitive explanation for its heightened performance. Eventually, our refined template denoising technique is expounded upon in subsection 3.3.

3.1   Chain-of-Thought-Styled Two-stage Sentence Representation

To effectively adapt CoT to new application scenarios, it is essential to clarify its main characteristics. Formally, CoT consists of two principal components: the reasoning process and the resultant conclusions. The former is a model-generated sequence predicated on given context, while the latter constitutes the final outcome derived from synthesizing the intermediate reasoning steps. Collectively, they form a problem-solving strategy that progresses from simple to complex.

CoT primarily operates on high-parameter generative PLMs. Although discriminative models like BERT possess robust natural language understanding capabilities, they cannot directly generate intermediate reasoning steps. Therefore, to guide such models towards CoT-style progressive reasoning, we need to employ prompt engineering to artificially design a multi-step sentence representation pipeline that is both versatile and adaptive.

Versatility here refers to the method's applicability to various sentences types, which requires a general strategy for solving sentence representation tasks. Adaptability, in contrast, emphasizes the model's ability to identify and prioritize key information from sentences based on their distinct semantics. Drawing inspiration from human practices in text summarization, we propose a two-stage sentence embedding derivation methodology: comprehension followed by summarization. Each stage incorporates a [MASK] token, enabling the model to attribute contextually relevant interpretations to [MASK] across diverse scenarios.

Following the aforementioned principles, our devised templates are detailed in Table 1. Given that the training set for unsupervised sentence representation tasks only includes a series of standalone sentences, different templates are adopted to construct anchor sentences, positive instances, and hard negative instances. These elements are subsequently integrated into our revised contrastive learning loss function for training, a procedure elaborated upon in subsection 3.2. After forward computation, we take the output vector corresponding to the final [MASK] token as the sentence embedding for the input text, mirroring the essence of CoT reasoning, which focuses on the model's conclusive output rather than the intermediate process.

In section 4, this paper will further substantiate the rationality of CoT-BERT's two-stage sentence representation method and its correlation with CoT

Table 1. Manual templates for our prompt-based learning. To alleviate burden and ensure conciseness in CoT-BERT, the templates selected for the anchor sentence, positive instance, and hard negative instance only have slight differences.

Anchor Sentence

The sentence of "[X]" means [MASK], so it can be summarized as [MASK].

Positive Instance

The sentence : "[X]" means [MASK], so it can be summarized as [MASK].

Hard Negative Instance

The sentence : "[X]" does not mean [MASK], so it cannot be summarized as [MASK].

---

through an array of experiments. Specifically, we aim to explore the following pivotal inquiries:

- Does the two-stage sentence representation strategy of CoT-BERT surpass the use of any single sub-stage alone?
- Does the progressive relationship between sub-stages contribute to enhanced model performance? If affirmative, can a genuine progressive linkage between comprehension and summarization be established?
- Compared to static templates, does the adaptive reasoning facilitated by the [MASK] token exhibit better universality?

Exploring these questions will deepen our understanding of CoT-inspired multi-stage reasoning. Empirical validation is critical for providing definitive answers to these considerations.

3.2   Enhancing Contrast with Extended InfoNCE Loss

Current methods for unsupervised sentence representation through contrastive learning commonly employ the InfoNCE Loss to guide the training process. For any given sentence x_i within a batch of size N, where $f(\cdot)$ denotes the encoding technique and τ represents a temperature coefficient, the InfoNCE Loss is defined as follows:

$$\ell_i = -\log\frac{e^{\text{sim}(f(x_i),f(x_i)^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(f(x_i),f(x_j)^+)/\tau}} \tag{1}$$

In Equation 1, sim signifies a similarity metric between two sentence vectors, typically cosine similarity. This formula encourages the model to maximize the similarity between the anchor sentence vector $f(x_i)$ and its positive instance $f(x_i)^+$, while simultaneously diminishing the similarity between $f(x_i)$ and other unrelated sentence vectors $f(x_j)$ within the same batch. Both Sim-CSE and PromptBERT adopt an InfoNCE Loss of this form. In SimCSE, positive instances are crafted using the intrinsic dropout of Transformer blocks,

whereas PromptBERT employs different templates to generate positive samples. Although PromptBERT's comparative experiments reveal minimal performance distinctions between these two methods under their manual templates, the utilization of varied templates for data augmentation can additionally facilitate the creation of hard negatives.

Specifically, ConPVP designs multiple templates and integrates the comparison between the anchor sentence vector $f(x_i)$ and the negative instance $f(x_j)^-$ in the denominator of Equation 1. By contrast, our CoT-BERT extends this by also introducing a comparison between the positive instance $f(x_i)^+$ and the negative sample $f(x_j)^-$. The differences among these approaches are visualized in Figure 1. Intuitively, our refined InfoNCE Loss incorporates more reference items into the sentence representation calculation process, rendering the distribution disparities among unrelated sentence vectors more pronounced within the semantic space. Subsection 4.4 will further elaborate on the effects of this modification with corresponding experimental results.
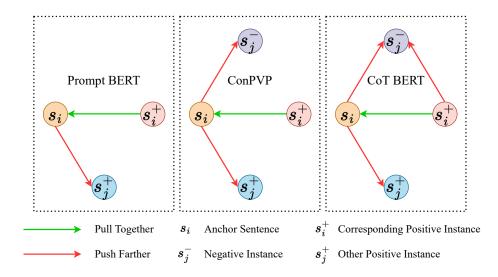

Fig. 1. Behavior of three variants of InfoNCE Loss within the semantic space of BERT. For clarity, we depict this figure with the anchor sentence s_i as the focal point.

Formalization of the described process leads to the ultimate loss function employed by CoT-BERT, as illustrated in Equation 2 and 3. Here, $\boldsymbol{r}_i$ represents the template-denoised sentence embedding for x_i. Particularly, in this context, the standard InfoNCE Loss can be expressed as $\ell_i = -\log\frac{a_{ii+}}{\sum_{j=1}^{N} a_{ij+}}$.

$$a_{ii+} = \exp\left(\frac{\cos(\boldsymbol{r}_i, \boldsymbol{r}_i^+)}{\tau}\right) \qquad a_{ij+} = \exp\left(\frac{\cos(\boldsymbol{r}_i, \boldsymbol{r}_j^+)}{\tau}\right)$$
$$a_{ij-} = \exp\left(\frac{\cos(\boldsymbol{r}_i, \boldsymbol{r}_j^-)}{\tau}\right) \qquad a_{i+j-} = \exp\left(\frac{\cos(\boldsymbol{r}_i^+, \boldsymbol{r}_j^-)}{\tau}\right) \tag{2}$$

$$\text{ours:} \quad \ell_i = -\log\frac{a_{ii+}}{\sum_{j=1}^{N}\left(a_{ij+} + a_{ij-} + a_{i+j-}\right)} \tag{3}$$

3.3   Leveraging [PAD] Tokens for Template Denoising

Deriving sentence representations through prompts involves the concatenation of templates with input sentences, which are then fed into the encoder for a complete forward pass. However, due to the attention mechanism, the output vector of the [MASK] token is influenced by the presence of the template, potentially distorting its original semantics.

To address this concern, PromptBERT suggests sending an empty template into BERT while maintaining position ids identical to those utilized when incorporating the input sentence. This process generates the template bias $\hat{\boldsymbol{h}}_i$. Ultimately, the embedding $\boldsymbol{r}_i$ used in the contrastive learning objective function is the difference between the sentence vector $\boldsymbol{h}_i$ and the template bias $\hat{\boldsymbol{h}}_i$.

From our perspective, a more effective approach than modeling input sentences with identical position ids is to populate [PAD] placeholders, matching the length of input sentences. This strategy naturally aligns position ids. Additionally, [PAD] placeholders can construct sentences devoid of significant meaning, thus serving to represent the inherent semantics of the template.

Figure 2 provides a detailed illustration of our template denoising method. In line with our two-stage representation derivation method, we extract the embedding corresponding to the final [MASK] token as the representation for the empty template. Since samples are processed in batches within the encoder, we pad each template to a predefined maximum length. Subsequently, we set the attention masks for the empty template to 1, while assigning 0 to the attention masks for the padded tail.
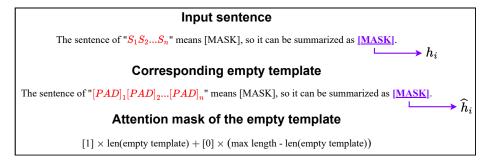

Input sentence

The sentence of "$S_1 S_2 ... S_n$" means [MASK], so it can be summarized as [MASK]. $\longrightarrow h_i$

Corresponding empty template

The sentence of "$[PAD]_1 [PAD]_2 ... [PAD]_n$" means [MASK], so it can be summarized as [MASK]. $\longrightarrow \widehat{h}_i$

Attention mask of the empty template

$[1] \times \text{len(empty template)} + [0] \times \big(\text{max length - len(empty template)}\big)$

Fig. 2. Illustration of the template denoising method employed by CoT-BERT. In this depiction, we utilize the template for anchor sentences as an example, with analogous treatment applied to both positive and hard negative instances.

The comparison between our designed denoising strategy and the method proposed by PromptBERT, conducted on both BERT and RoBERTa, is discussed in subsection 4.5.

4 Experiment

4.1 Setup

In our experimental setup, we adhere to the established conventions by utilizing the SentEval [24] toolkit for the evaluation of our model, CoT-BERT, across seven distinct English STS tasks [25–31], with Spearman correlation as the metric.

Our training dataset is sourced from SimCSE [11], comprising 10^6 sentences randomly sampled from English Wikipedia. During the training phase, we save checkpoints based on our model's performance on the development set of STS-Benchmark. In the validation phase, we directly compute the cosine similarity between sentence embeddings without any additional regressors. Furthermore, akin to PromptBERT, we only perform template denoising during the training process.

Our selection of baselines includes a diverse range of models, encompassing both non-BERT-based methods such as GloVe [32] and USE [33], as well as the latest advancements in unsupervised sentence representation built upon BERT architecture. Specifically, the models chosen for comparison include BERT-flow, BERT-whitening, IS-BERT [34], ConSERT, SimCSE, DCLR, ArcCSE, ESim-CSE, DiffCSE, PCL, PromCSE, PromptBERT, ConPVP, RankEncoder, and RankCSE. This set of baselines serves to thoroughly validate the effectiveness of CoT-BERT from various perspectives.

4.2 Main Results

Table 2 provides a comprehensive overview of our primary experimental results across multiple STS datasets. When $\text{RoBERTa}_{\text{base}}$ is employed as the encoder, CoT-BERT achieves an exceptional Spearman's correlation score of 80.62, setting a new record for the current best performance. In scenarios where $\text{BERT}_{\text{base}}$ serves as the encoder, our model attains an average Spearman correlation of 79.40 on the seven STS tasks. While this value falls slightly below the existing SOTA results achieved by RankCSE and RankEncoder, it is imperative to note that CoT-BERT's architecture exclusively relies on BERT itself, whereas RankCSE and RankEncoder incorporate other text representation models such as SimCSE, SNCSE [35], or external database. Therefore, CoT-BERT's performance remains highly competitive and is obtained with a concise model configuration.

Moreover, compared to SimCSE, CoT-BERT outperforms it by 3.15% and 4.05% on $\text{BERT}_{\text{base}}$ and $\text{RoBERTa}_{\text{base}}$, respectively. In contrast to Prompt-BERT, which also leverages prompt-based techniques, CoT-BERT exhibits a 0.86% improvement on $\text{BERT}_{\text{base}}$ and a 1.47% enhancement on $\text{RoBERTa}_{\text{base}}$. These results collectively provide compelling evidence attesting to the effectiveness of CoT-BERT.

Table 2. Performance of different models on STS tasks under **unsupervised** settings. Consistent with established research conventions, the table displays the Spearman's rank correlation between model predictions and human-annotated scores. Baseline results are sourced from original papers. The highest scores obtained by models using the same PLM for each dataset are highlighted in bold.

PLMs	Methods	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Non-BERT	GloVe(avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
	USE	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
BERT$_{base}$	BERT-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
	BERT-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
	IS-BERT	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
	ConSERT	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
	SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
	DCLR	70.81	83.73	75.11	82.56	78.44	78.31	71.59	77.22
	ArcCSE	72.08	84.27	76.25	82.32	79.54	79.92	72.39	78.11
	ESimCSE	73.40	83.27	77.25	82.66	78.81	80.17	72.30	78.27
	DiffCSE	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
	PCL	72.84	83.81	76.52	83.06	79.32	80.01	73.38	78.42
	PromCSE	73.03	85.18	76.70	84.19	79.69	80.62	70.00	78.49
	PromptBERT	71.56	84.58	76.98	84.47	80.60	81.60	69.87	78.54
	ConPVP	71.72	84.95	77.68	83.64	79.76	80.82	73.38	78.85
	RankCSE$_{listNet}$	74.38	**85.97**	77.51	84.46	**81.31**	81.46	75.26	80.05
	RankEncoder	**74.88**	85.59	**78.61**	83.50	80.56	81.55	**75.78**	**80.07**
	CoT-BERT	72.56	85.53	77.91	**85.05**	80.94	**82.40**	71.41	79.40
RoBERTa$_{base}$	SimCSE	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
	DCLR	70.01	83.08	75.09	83.66	81.06	81.86	70.33	77.87
	DiffCSE	70.05	83.43	75.49	82.81	82.12	82.38	71.19	78.21
	PCL	71.13	82.38	75.40	83.07	81.98	81.63	69.72	77.90
	ESimCSE	69.90	82.50	74.68	83.19	80.30	80.99	70.54	77.44
	PromptRoBERTa	73.94	84.74	77.28	84.99	81.74	81.88	69.50	79.15
	ConPVP	73.20	83.22	76.24	83.37	81.49	82.18	**74.59**	79.18
	RankCSE$_{listNet}$	72.91	**85.72**	76.94	84.52	**82.59**	**83.46**	71.94	79.73
	CoT-RoBERTa	**75.43**	85.47	**78.74**	**85.64**	82.21	83.40	73.46	**80.62**

4.3 Assessing our Two-stage Sentence Representation

To further ascertain the efficacy of CoT-BERT's two-stage sentence representation strategy and address the inquiries raised in subsection 3.1, we embark on an extensive series of experiments. For the manual templates provided in Table 1, we define the initial and latter parts of the template as its *prefix* and *suffix*, respectively. To illustrate, for the anchor sentence template: "The sentence of '[X]' means [MASK], so it can be summarized as [MASK].", we designate its *prefix* as "The sentence of '[X]' means [MASK]." and its *suffix* as "The sentence of '[X]' can be summarized as [MASK]." When we refer to "*prefix + suffix*", it implies the use of our complete, original templates. The handling of positive instances and hard negative instances follows a similar pattern.

Multi-stage vs. Single-stage: Here, we explore whether CoT-BERT's two-stage sentence representation method surpasses the performance of each sub-stage when deployed independently. To this end, we conduct two sets of com-

parative experiments: *prefix* vs. *prefix + suffix* and *suffix* vs. *prefix + suffix*. If *prefix + suffix* consistently performs better, it indicates that the integrative application of both stages effectively bolsters the model's performance.

Progressive Relationship Between Sub-stages: In this part, we replace CoT-BERT's original template *prefix* with a non-sequitur (termed *irrelevant prefix*) and examine whether the performance of *irrelevant prefix + suffix* deteriorates compared to original *prefix + suffix*, thereby proving the necessity of a logical connection between the two sub-stages. An example modification includes changing the anchor sentence template to "Penguin is a flightless bird, and the sentence of '[X]' can be summarized as [MASK]." The positive and hard negative instance templates are treated in the same manner.

Additionally, by reversing the order of the two sub-stages in CoT-BERT templates to *suffix + prefix*, we seek to demonstrate the significance of the progressive reasoning process from comprehension to summarization. For instance, the anchor sentence template is converted to "The sentence of '[X]' can be summarized as [MASK], so it means [MASK]." The positive instance template and hard negative instance template are adjusted accordingly.

Adaptivity in Multi-stage Reasoning: Herein, we substitute the first [MASK] token in CoT-BERT templates with a static element (*static prefix + suffix*) to observe performance disparities against the dynamic *prefix + suffix* setup. This experiment aims to underscore the pivotal role of adaptive reasoning facilitated by the intermediate [MASK] token, notwithstanding that the derivation of sentence embeddings is from the final [MASK] token. For illustration, the anchor sentence template is changed to "The sentence of '[X]' means something, so it can be summarized as [MASK].", with corresponding adjustments made to the templates for positive and hard negative instances.

The results of these experiments are recorded in Table 3, demonstrating that the *prefix + suffix* configuration achieves the highest average Spearman's correlation score across seven STS benchmarks. These findings affirm the rationality and effectiveness of CoT-BERT's two-stage sentence representation strategy. Additionally, the specific settings and sequencing of the comprehension and summarization stages, together with the introduction of adaptive [MASK] tokens, adeptly emulate the essence of CoT reasoning.

Table 3. Comparative analysis of CoT-BERT's two-stage sentence representation.

	$BERT_{base}$
only prefix	78.91
only suffix	78.83
irrelevant prefix + suffix	78.44
static prefix + suffix	78.91
suffix + prefix	78.46
prefix + suffix	**79.40**

4.4   Evaluating our Extended InfoNCE Loss

In subsection 3.2, we have expounded upon our proposed Extended InfoNCE Loss and its design concept. Departing from prior work, we introduce a comparison between the anchor sentence and negative samples in the loss function's denominator, as well as a comparison between the positive and negative instances, with the latter being proposed by CoT-BERT for the first time. In line with the symbol definitions in Equation 2, we maintain the usage of a_{i+j-} to denote the similarity computation between the positive and negative instances.

Table 4 displays the performance disparities exhibited by CoT-BERT across seven STS tasks, both with and without the introduction of a_{i+j-}. The values in the table correspond to the average Spearman correlation. Remarkably, irrespective of whether $BERT_{base}$ or $RoBERTa_{base}$ serves as the PLM, our extended InfoNCE Loss consistently yields improvements in the model's performance.

Table 4. Ablation experiments on the extended InfoNCE Loss for CoT-BERT.

	$BERT_{base}$	$RoBERTa_{base}$
CoT-BERT (without a_{i+j-})	79.11	80.30
CoT-BERT (with a_{i+j-})	**79.40**	**80.62**

To offer deeper insights into the underlying factors underpinning these results, we conduct alignment and uniformity analyses on three models: Prompt-BERT, CoT-BERT (with and without a_{i+j-}). This evaluation utilizes the STS-B test set, a corpus encompassing a total of 1,379 sentence pairs, each accompanied by a similarity score ranging from 0.0 to 5.0.

For the computation of uniformity, the entire set of 1,379 sentence pairs is employed. In the case of alignment calculation, we filter for sentence pairs with similarity scores greater than 4.0. The results are detailed in Table 5.

Both alignment and uniformity serve as established metrics for evaluating the model's semantic space quality. Given a data distribution p_{data}, alignment quantifies the expected distance between samples and their corresponding positive instances, defined as below:

$$\ell_{align} \triangleq \mathop{\mathbb{E}}_{(x,x^+) \sim p_{data}} \|f(x) - f(x^+)\|^2 \tag{4}$$

Uniformity, on the other hand, reflects the overall evenness of the sentence vector space by calculating the average distance between embeddings of any two semantically unrelated texts:

$$\ell_{uniform} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{data}} e^{-2\|f(x)-f(y)\|^2} \tag{5}$$

The experimental outcomes in Table 5 provide empirical evidence in support of the intuitive explanations outlined in subsection 3.2. We posit that the incorporation of a_{ij-} introduces additional contextual references during the training

process, thus enhancing the model's discriminative power among diverse samples. The inclusion of a_{i+j-} further amplifies this effect by concurrently distinguishing negative instances from both anchor sentences and positive instances, thereby leading to an augmentation in the model's uniformity.

Table 5. Results of alignment and uniformity calculations performed on models using the STS-B test set. Lower values indicate better performance. All three sets of experiments employed RoBERTa$_{base}$ as the pre-trained model.

PLM = RoBERTa$_{base}$	Alignment	Uniformity
PromptBERT	0.0957	- 1.2033
CoT-BERT (without a_{i+j-})	0.1089	- 1.3852
CoT-BERT (with a_{i+j-})	0.1278	- 1.5492

Meanwhile, a diminution in the alignment metric is observed. We attribute this phenomenon to two potential factors. Firstly, while computing alignment with the STS-B test set, we categorize sentence pairs with a similarity exceeding 4.0 as positive examples, a threshold selection that may introduce bias. Additionally, when modifying the objective function to impose more stringent requirements on the model, there is no corresponding increase in the number of model iterations. Consequently, within a constrained number of update steps, the model may struggle to optimally adjust its alignment.

4.5   Assessing our Template Denoising Strategy

We also embark on an ablation study regarding the template denoising method of CoT-BERT. To distinguish between the two denoising strategies under examination, we refer to the technique introduced by PromptBERT as "position denoise" and our proposed approach as "[PAD] denoise."

The primary difference between these two lies in the fact that "[PAD] denoise" does not deliberately adjust the values of position ids but rather models an empty template by injecting [PAD] placeholders of identical length as the input sentence, accompanied by corresponding adaptations to the attention masks.

Our evaluation of different denoising methods on the seven STS tasks is presented in Table 6. As in our prior assessments, we continue to report the model's average Spearman correlation.

The empirical findings compellingly demonstrate the advantages conferred by the introduction of denoising encodings within the contrastive learning loss function. Moreover, "[PAD] denoise" notably outperforms its "position denoise" counterpart.

One possible explanation for the effectiveness of template denoising is that sentence embeddings derived from PLMs may contain some general information, such as syntax and sentence structure. During the contrastive learning process, subtracting that kind of information helps highlight the distinctions between input sentences, thereby fostering more precise clustering outcomes.

Table 6. Ablation study of CoT-BERT's template denoising strategy.

	BERT$_{base}$	RoBERTa$_{base}$
CoT-BERT (without denoise)	78.69	79.87
CoT-BERT (position denoise)	78.89	79.95
CoT-BERT ([PAD] denoise)	**79.40**	**80.62**

5   Discussion

5.1   Distribution of Predicted Values

We present a visualization of the predicted value distribution generated by CoT-BERT using the STS-B test set in Figure 3. It should be noted that the model remains unexposed to these data during its training phase. Therefore, the model's performance on this dataset can serve as a reliable indicator of its overall efficacy.
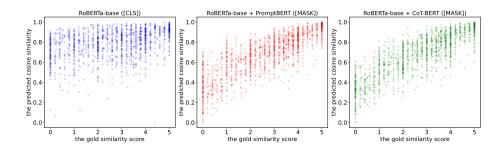

Fig. 3. Correlation diagram between the true similarity scores and model-predicted cosine similarity on the STS-B test set. The vertical axis has been normalized for clarity, and the methods employed for deriving sentence embeddings ([CLS] or [MASK]) are explicitly indicated for reference.

As illustrated in Figure 3, the initial RoBERTa checkpoint demonstrates limited discriminative ability for sentence pairs with varying degrees of similarity. It tends to yield higher predicted values across the board. In contrast, both PromptBERT and CoT-BERT display a clear upward trend in predicted values as the similarity between sentences increases.

Furthermore, CoT-BERT distinctly outperforms PromptBERT, especially in handling samples with annotated similarity scores ranging from 0 to 2. CoT-BERT's predictions within this range are considerably more concentrated, indicating enhanced precision. Additionally, it is noteworthy that for some samples with a true similarity score of 1, PromptBERT's predicted values are even higher than those for most samples with a true similarity score of 2, while CoT-BERT does not exhibit such a pattern.

5.2   Introducing More Stages

In subsection 3.1, we have elucidated the underlying design principles behind CoT-BERT's two-stage sentence representation method. Furthermore, in subsection 4.3, we have empirically demonstrated the superiority of this approach over using either sub-stage in isolation. Naturally, this leads to a pertinent question: if we further divide the template and introduce more sub-stages, will the model's performance continue to improve?

Regrettably, due to constraints on computational resources, we are currently unable to conduct experiments in this regard. Nevertheless, we recognize several crucial considerations that warrant careful attention in such investigations.

Firstly, increasing the number of stages within the template inherently augments the complexity of the corresponding prompt. This heightened complexity demands substantial effort in devising and selecting the most suitable prompts. Additionally, even if a template performs well on a specific PLM, its adaptability to other PLMs remains uncertain. Besides, as the template's length expands, the weight of the input sentence [X] within the prompt gradually diminishes, and its distance from the final [MASK] token increases. This could potentially result in the model inadequately capturing the semantics of [X]. Moreover, it's worth mentioning that certain concise short sentences in natural language may not lend themselves well to being segmented into multiple stages. Lastly, the presence of templates compresses the maximum input length that a PLM can accommodate, thereby affecting the model's capacity to handle longer texts. We leave the exploration of these aspects to future work.

6   Conclusion

In this study, we propose CoT-BERT, a pioneering strategy for sentence representation computation. To the best of our knowledge, CoT-BERT is the first work that combines the Chain-of-Thought (CoT) concept with text representation tasks. Furthermore, we improve the distribution of sentence embeddings within the BERT semantic space by introducing an extended InfoNCE Loss. Additionally, we devise a more efficacious template denoising method to mitigate the impact of prompt-induced biases on sentence semantics.

Experimental findings across seven Semantic Textual Similarity tasks unequivocally affirm the outstanding efficacy of CoT-BERT. It surpasses a spectrum of formidable baselines and achieves state-of-the-art performance without relying on additional text representation models or external databases. Comprehensive ablation experiments demonstrate that CoT-BERT's two-stage sentence representation, extended InfoNCE loss, and refined template denoising methods collectively contribute to the enhancements in its overall performance.

References

1. Asai, A., Min, S., Zhong, Z., Chen, D.: Retrieval-based Language Models and Applications. In: Proceedings of the 61st Annual Meeting of the Association for Com-

putational Linguistics (Volume 6: Tutorial Abstracts), pp. 41–46. (2023). `https://doi.org/10.18653/v1/2023.acl-tutorials.6`

2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. (2019). `https://doi.org/10.18653/v1/N19-1423`

3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)

4. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the Sentence Embeddings from Pre-trained Language Models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9119–9130. (2020). `https://doi.org/10.18653/v1/2020.emnlp-main.733`

5. Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., Wei, F., Huang, H., Deng, D., Zhang, Q.: PromptBERT: Improving BERT Sentence Embeddings with Prompts. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 8826–8837. (2022). `https://doi.org/10.18653/v1/2022.emnlp-main.603`

6. Zeng, J., Yin, Y., Jiang, Y., Wu, S., Cao, Y.: Contrastive Learning with Prompt-derived Virtual Semantic Prototypes for Unsupervised Sentence Embedding. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 7042–7053. (2022). `https://doi.org/10.18653/v1/2022.findings-emnlp.522`

7. Chuang, Y.-S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljacic, M., Li, S.-W., Yih, S., Kim, Y., Glass, J.: DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4207–4218. (2022). `https://doi.org/10.18653/v1/2022.naacl-main.311`

8. Liu, J., Liu, J., Wang, Q., Wang, J., Wu, W., Xian, Y., Zhao, D., Chen, K., Yan, R.: RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 13785–13802. (2023). `https://doi.org/10.18653/v1/2023.acl-long.771`

9. Seonwoo, Y., Wang, G., Seo, C., Choudhary, S., Li, J., Li, X., Xu, P., Park, S., Oh, A.: Ranking-Enhanced Unsupervised Sentence Representation Learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 15783–15798. (2023). `https://doi.org/10.18653/v1/2023.acl-long.879`

10. Wang, X., He, J., Wang, P., Zhou, Y., Sun, T., Qiu, X.: DenoSent: A Denoising Objective for Self-Supervised Sentence Representation Learning. arXiv preprint arXiv:2401.13621 (2024)

11. Gao, T., Yao, X., Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910. (2021). `https://doi.org/10.18653/v1/2021.emnlp-main.552`

12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)

13. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
15. Oord, A. van den, Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2018)
16. Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., Xu, W.: ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 5065–5075. (2021). `https://doi.org/10.18653/v1/2021.acl-long.393`
17. Zhang, Y., Zhu, H., Wang, Y., Xu, N., Li, X., Zhao, B.: A Contrastive Framework for Learning Sentence Representations from Pairwise and Triple-wise Perspective in Angular Space. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 4892–4903. (2022). `https://doi.org/10.18653/v1/2022.acl-long.336`
18. Wu, X., Gao, C., Zang, L., Han, J., Wang, Z., Hu, S.: ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3898–3907. (2022).
19. Zhou, K., Zhang, B., Zhao, X., Wen, J.-R.: Debiased Contrastive Learning of Un-supervised Sentence Representations. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 6120–6130. (2022).
20. Wu, Q., Tao, C., Shen, T., Xu, C., Geng, X., Jiang, D.: PCL: Peer-Contrastive Learning with Diverse Augmentations for Unsupervised Sentence Embeddings. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 12052–12066. (2022).
21. Su, J., Cao, J., Liu, W., Ou, Y.: Whitening sentence representations for better semantics and faster retrieval. arXiv preprint arXiv:2103.15316 (2021)
22. Chen, N., Shou, L., Pei, J., Gong, M., Cao, B., Chang, J., Li, J., Jiang, D.: Allevi-ating Over-smoothing for Unsupervised Sentence Representation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3552–3566. (2023).
23. Jiang, Y., Zhang, L., Wang, W.: Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 3021–3035. (2022). `https://doi.org/10.18653/v1/2022.findings-emnlp.220`
24. Conneau, A., Kiela, D.: SentEval: An Evaluation Toolkit for Universal Sentence Representations. In: Proceedings of the Eleventh International Conference on Lan-guage Resources and Evaluation (LREC 2018). (2018).
25. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 385–393. (2012).
26. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: *SEM 2013 shared task: Semantic Textual Similarity. In: Second Joint Conference on Lexical and Com-putational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pp. 32–43. (2013).
27. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2014 Task 10: Multilingual Semantic

Textual Similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 81–91. (2014). `https://doi.org/10.3115/v1/S14-2010`.

28. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 252–263. (2015). `https://doi.org/10.18653/v1/S15-2045`.

29. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 497–511. (2016). `https://doi.org/10.18653/v1/S16-1081`.

30. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 1–14. (2017). `https://doi.org/10.18653/v1/S17-2001`.

31. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 216–223. (2014).

32. Pennington, J., Socher, R., Manning, C.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. (2014). `https://doi.org/10.3115/v1/D14-1162`.

33. Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal Sentence Encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 169–174. (2018). `https://doi.org/10.18653/v1/D18-2029`.

34. Zhang, Y., He, R., Liu, Z., Lim, K.H., Bing, L.: An Unsupervised Sentence Embedding Method by Mutual Information Maximization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1601–1610. (2020). `https://doi.org/10.18653/v1/2020.emnlp-main.124`.

35. Wang, H., Dou, Y.: SNCSE: contrastive learning for unsupervised sentence embedding with soft negative samples. In: International Conference on Intelligent Computing, pp. 419–431. Springer (2023)