Adaptive Reinforcement Tuning Language Models as Hard Data Generators for Sentence Representation

Bo Xu, Yifei Wu, Shouang Wei, Ming Du, Hongya Wang
School of Computer Science and Technology, Donghua University, Shanghai, China
xubo@dhu.edu.cn, wunein@mail.dhu.edu.cn, 2212538@mail.dhu.edu.cn,
duming@dhu.edu.cn, hywang@dhu.edu.cn

Abstract

Sentence representation learning is a fundamental task in NLP. Existing methods use contrastive learning (CL) to learn effective sentence representations, which benefit from high-quality contrastive data but require extensive human annotation. Large language models (LLMs) like ChatGPT and GPT4 can automatically generate such data. However, this alternative strategy also encounters challenges: 1) obtaining high-quality generated data from small-parameter LLMs is difficult, and 2) inefficient utilization of the generated data. To address these challenges, we propose a novel adaptive reinforcement tuning (ART) framework. Specifically, to address the first challenge, we introduce a reinforcement learning approach for fine-tuning small-parameter LLMs, enabling the generation of high-quality hard contrastive data without human feedback. To address the second challenge, we propose an adaptive iterative framework to guide the small-parameter LLMs to generate progressively harder samples through multiple iterations, thereby maximizing the utility of generated data. Experiments conducted on seven semantic text similarity tasks demonstrate that the sentence representation models trained using the synthetic data generated by our proposed method achieve state-of-the-art performance. Our code is available at https://github.com/WuNein/AdaptCL.

Keywords: sentence representation, contrastive learning, data augmentation, reinforcement learning

1. Introduction

Sentence representation learning is a fundamental task in natural language processing (NLP), which encodes sentences into fixed-dimensional vector representations that capture their semantic meaning or contextual information. These representations serve as a foundation for various downstream NLP tasks such as text classification (Suresh and Ong, 2021), question answering (Karpukhin et al., 2020) and machine translation (Pan et al., 2021).

Existing sentence representation methods primarily leverage the *contrastive learning* (CL) paradigm to learn effective sentence representations, and the use of high-quality contrastive data can significantly enhance the performance of sentence representation models. The principle of contrastive learning involves guiding sentence representation models to differentiate between positive (similar) and negative (dissimilar) sentences. There are several approaches to obtain contrastive data, including various data augmentation methods (e.g., *dropout* (Gao et al., 2021b), *word repetition* (Wu et al., 2022b), *case switch* and *retrieved negative* (Wang et al., 2022c)) and direct utilization of existing human-annotated datasets (e.g., *QQP*, *Flicker30K*, *ParaNMT* and *NLI* (Gao et al., 2021b)). Among these approaches, direct utilization of human-annotated *NLI* data has yielded the most favorable results. This highlights the importance of high-quality contrastive data, but annotating such data demands substantial human effort. Therefore, a natural question arises: *can we explore the possibility of automatically generating high-quality contrastive data to further improve the performance of sentence representation models*?

Recently, large language models (LLMs), such as ChatGPT (Ouyang et al., 2022) and GPT4 (OpenAI, 2023), have achieved success as data generators in various NLP tasks (Zhang et al., 2023). This achievement has opened up new possibilities for automatically generating high-quality synthetic contrastive data. However, existing methods heavily rely on large-parameter LLMs [1] for direct data generation, resulting in significant API expenses or high costs for local deployment (Hsieh et al., 2023). Even when utilizing medium-parameter LLMs, the expenses remain substantial.

A more cost-effective strategy is to utilize small-parameter LLMs to generate synthetic data and enhance sentence representation models. However, this alternative strategy also encounters challenges:

- Firstly, obtaining high-quality data from small-parameter LLMs is difficult. An intuitive approach is to directly generate data from the small-parameter LLMs, However, the data quality is poor: 1) the generated data may contain errors. Requesting a positive sample might result in a negative one, and vice versa.

[1] In this paper, we roughly categorize these large language models by their model sizes: *small-parameter* LLMs (less than 20B, such as WizardLM 7B (Xu et al., 2023)), *medium-parameter* LLMs (between 20B and 100B, such as LLAMA 70B (Touvron et al., 2023)) and *large-parameter* LLMs (over 100B, such as ChatGPT and GPT4).

LREC-COLING 2024, pages 358–371
20-25 May, 2024.

2) The generated data might not be sufficiently hard, leading to minimal improvements in the performance of existing sentence representation models. An alternative approach is to use the fine-tuned small-parameter LLMs to generate data, but lack of supervised data.

- Secondly, the methods for utilizing generated data are inefficient. Existing methods typically generate a large amount of homogeneous data at once to ensure adequate model training. However, only a small portion of this data is eventually utilized. This approach not only results in a waste of computational resources, but also fails to fully utilize the data generation capabilities of language models.

To address these challenges, we propose a novel adaptive reinforcement tuning (ART) framework. This framework aims to optimize small-parameter LLMs in a reinforcement learning manner to adaptively generate hard contrastive samples with multiple iterations. These samples are then used to enhance the performance of existing sentence representation models. Specifically, to address the first challenge, we introduce a reinforcement learning approach for fine-tuning small-parameter LLMs, enabling the generation of high-quality contrastive data without human feedback. The key to this reinforcement learning approach lies in the creation of a novel dual reward model, capable of automatically assessing the correctness and difficulty of the data generated by the LLMs. To address the second challenge, we propose an adaptive iterative framework to guide small-parameter LLMs to generate progressively harder samples through multiple iterations, thereby maximizing the utility of generated sentence data.

Our main contributions can be summarized as follows:

- Firstly, we propose a more cost-effective strategy to leverage large language models for automatically generating high-quality synthetic contrastive data to enhance the performance of sentence representation models. This strategy employs small-parameter LLMs, yet achieves results comparable to or even surpassing those obtained by directly using large-parameter LLMs.
- Secondly, we propose a novel adaptive reinforcement tuning framework to optimize small-parameter LLMs in a reinforcement learning manner. Through multiple iterations, this framework adaptively generates hard contrastive samples, thereby maximizing the utility of the generated data.
- Finally, experiments conducted on seven semantic text similarity tasks demonstrate that the sentence representation models trained using the synthetic data generated by our proposed method achieve state-of-the-art performance. We also conducted ablation studies to showcase the critical roles played by the reinforcement learning approach and the adaptive iterative framework within our proposed framework.

2. Overview

In this section, we first formulate our problem, and then introduce the framework of our system: adaptive reinforcement tuning (ART).

2.1. Problem Formulation

Given the supervised training corpus $\mathcal{X}$, unsupervised corpus D and the existing supervised sentence encoder f_θ, where $\mathcal{X}$ consists of a set of labeled contrastive data $\{x_i, x_i^+, x_i^-\}$, D consists of a set of unlabeled sentences and the supervised sentence encoder f_θ is initially trained with $\mathcal{X}$. Our objective is to generate synthetic contrastive sentence sample data $\tilde{\mathcal{X}}$ from unsupervised corpus D to improve the performance of existing supervised sentence encoder f_θ.

2.2. Framework

Our adaptive reinforcement tuning (ART) framework for enhancing sentence representation is shown in Figure 1, consisting of an initialization backbone model training step followed by two adaptive iterative steps, namely the LLM fine-tuning step and the sentence encoder training step.

Specifically, the backbone model training step utilizes the NLI dataset to train an initial supervised sentence encoder and an NLI discriminator. The supervised sentence encoder serves as the model targeted for optimization in this paper. Additionally, both models play crucial roles in the subsequent adaptive iteration steps. The LLM fine-tuning step employs a reinforcement learning approach to fine-tune the large language model. This involves a dual reward model incorporating both the NLI discriminator and a difficulty evaluator based on the sentence encoder. The former is used to determine the correctness of the generated data, while the latter assesses the difficulty of the generated data. By combining the outputs of these two models, the final reward score is determined. The Proximal Policy Optimization (PPO) algorithm is applied to update the parameters of the large language model. The sentence encoder training step uses the synthetic data generated by the fine-tuned large language model to further train the sentence encoder. To ensure data quality, the NLI discriminator is employed to filter out the incorrectly generated sentences.

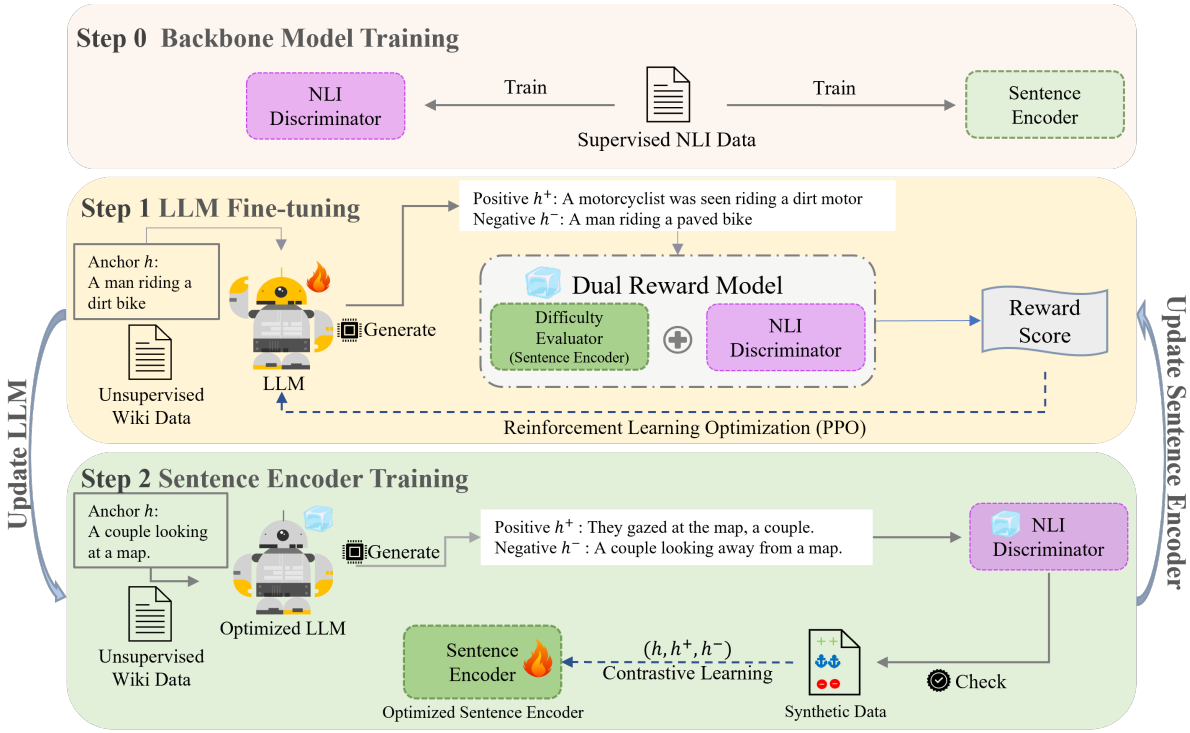

Figure 1: Our adaptive reinforcement tuning (ART) framework for enhancing sentence representation.

The two adaptive iteration steps can be executed in multiple rounds to continuously improve the performance of the sentence encoder.

3. Method

In this section, we introduce our framework in detail. The framework consists of an initialization backbone model training step followed by two adaptive iterative steps, namely the LLM fine-tuning step and the sentence encoder training step.

3.1. Backbone Model Training Step

The backbone model training step utilizes the NLI dataset to train an initial supervised sentence encoder and an NLI discriminator.

Since our objective is to leverage synthesized data to enhance the performance of existing sentence representation models rather than propose a new one, we directly employ existing sentence encoders, such as SimCSE (Gao et al., 2021b) or GenSE (Chen et al., 2022b).

We also employ the existing NLI models as our NLI discriminator, which is a three-class neural network model that includes a pre-trained model (e.g., RoBERTa-large) and a multi-layer perceptron (MLP) layer with a hidden layer. The initial purpose of the NLI discriminator is to determine whether a given pair of sentences entail, contradict, or are neutral to each other. In this paper, we leverage it to assess whether the positive and negative samples generated by the small-parameter LLM for a given sentence are correct. This assessment is based on whether the anchor sentence and the positive sample are *entailment*, and whether the anchor sentence and the negative sample are *contradiction*.

3.2. LLM Fine-tuning Step

The LLM fine-tuning step aims to improve the ability of the small-parameter LLM to generate hard synthetic data from unlabeled sentences D. Due to the lack of supervised data, we use reinforcement learning to fine-tune the small-parameter LLM. The LLM fine-tuning step involves a proximal policy optimization (PPO) algorithm and a dual reward model.

Positive Prompt: Generate a positive variation of Original Sentence, ensuring it has same meaning, exhibits different syntactical and grammatical structures. Original: "**[X]**" Positive:
Negative Prompt: Generate a negative variation of Original Sentence, ensuring it has a completely different meaning, similar syntax and grammar. Original: "**[X]**" Negative:

Table 1: Prompts used to generate hard positive and negative samples, respectively. [X] refers to the input (anchor) sentence.

3.2.1. Reinforcement Learning Optimization

We employ the proximal policy optimization (PPO) algorithm (Ouyang et al., 2022) to optimize the LLM on our environment to generate harder samples. The environment is a bandit environment that presents a positive/negative prompt and an anchor sample x from the unlabeled training corpora $\mathcal{D}$ and expects a positive/negative sample y to the corresponding prompts. The prompts used to generate hard positive and negative samples are shown in Table 1. Given the prompts and anchor-positive/negative sample pairs, it produces a reward determined by the reward model and ends the episode. In addition, we add a per-token KL penalty from the reference model, which is the frozen LLM before fine-tuning, at each token to mitigate over-optimization of the reward model. The objective function is defined as follows:

$$E_{(x,y)\sim\mathcal{D}}[r_\theta(x,y) - \beta D_{KL}(\pi_\phi^{\mathrm{RL}}(y \mid x)||\pi^{\mathrm{ref}}(y \mid x))], \tag{1}$$

where π_ϕ^{RL} is the learned RL policy, $r_\theta(x,y)$ is the reward score for output y with given input x, π^{ref} is the reference model, and β is the KL penalty coefficient. By fine-tuning LLMs with LoRA (Hu et al., 2021), we only need to keep one LLM in GPU memory. We switch between the reference model and the trained model by toggling LoRA components.

3.2.2. Dual Reward Model

Our dual reward model takes the input anchor x and the generated positive/negative sample y, and outputs a proper score $r_\theta(x,y)$. In this paper, we introduce a novel dual-reward model to assess the quality of samples generated by the LLM, which consists of an NLI discriminator and a difficulty evaluator based on the sentence encoder. The former is used to determine the correctness of the generated data, while the latter assesses the difficulty of the generated data.

Specifically, the NLI discriminator takes the (x,y) pair as input and outputs the probabilities of three labels, namely *entailment*, *contradiction* and *neutral*. Our expectation is that the sentences generated using positive prompts are in an *entailment* relationship with the anchor sentences, while the sentences generated using negative prompts are in a *contradiction* relationship with the anchor sentences. Therefore, we use score r_1 to represent the correctness of the generated data, which is defined as follows:

$$r_1 = P(t|x,y) - \Omega, \tag{2}$$

where $P(t|x,y)$ is the probability of the target label for the (x,y) pair, Ω is the lower bound of target probability.

The difficulty evaluator takes the (x,y) pair as input and outputs the difficulty of the pair. In our context, the hard positive samples of a given sentence are those with less similar text content while still preserving the same underlying semantic meaning, while hard negative samples are those that exhibit high text similarity but possess significantly different semantic meanings. Therefore, we use score r_2 to represent the difficulty of the pair (x,y), which is defined as follows:

$$r_2 = \begin{cases} (1 - \mathrm{sim}(x,y^+)) \cdot \mathrm{sgn}(\mathrm{sim}(x,y^+) - \alpha^+) \\ \mathrm{sim}(x,y^-) \cdot \mathrm{sgn}(\alpha^- - \mathrm{sim}(x,y^-)), \end{cases} \tag{3}$$

where $\mathrm{sim}(x,y^+)$ and $\mathrm{sim}(x,y^-)$ are the cosine similarity between anchor sample x and the positive/negative sample calculated by the sentence encoder, sgn is Sign function and α^+ and α^- are the lower bound and upper bounds for filter noisy positive and negative samples.

Finally, we combine both scores obtained from the NLI discriminator and the difficulty evaluator as the final reward score, which is defined as follows:

$$r_\theta(x,y) = w_1 \times r_1 + w_2 \times r_2, \tag{4}$$

where w_1 and w_2 are the weights for the NLI discriminator and the difficulty evaluator, respectively. The PPO algorithm is applied to update the parameters of the large language model.

3.3. Sentence Encoder Training Step

The sentence encoder training step uses the synthetic data generated by the fine-tuned small-parameter LLM to further train the sentence encoder, which consists of a data synthesis process and an encoder training process. To ensure data quality, the NLI discriminator is employed to filter out the incorrectly generated sentences.

Following existing works (Gao et al., 2021b; Chen et al., 2022b), we use the supervised SimCSE loss as the objective function, which is defined as follows:

$$\mathcal{L} = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}\left(e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau} + e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^-)/\tau}\right)}, \tag{5}$$

where $\mathbf{h}_i$, $\mathbf{h}_i^+$, and $\mathbf{h}_i^-$ represent the representations of the anchor, positive, and negative samples, respectively, and τ is a temperature hyper-parameter. Through training with high-quality synthesized samples, the sentence encoder is able to learn a better sentence representation.

4. Experiment

4.1. Experimental Setup

4.1.1. Data and Metrics

The **training data** is formatted as triplets, including an anchor, a positive sample, and a negative sample. The data is sourced from two categories: 1) **supervised NLI dataset**, which is a combination of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018a) datasets. 2) **synthetic data**, generated by our large language model through multi-round generation based on sentences from English Wikipedia. At each round, we generated 20k triplets. The **development data** is from the development set of Semantic Textual Similarities (STS) Benchmark (Cer et al., 2017). The **testing data** consists of seven standard STS datasets, including STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014) datasets. Except for the synthetic data, all other datasets were provided by SimCSE(Gao et al., 2021b).

Following previous methods (Gao et al., 2021b; Chen et al., 2022b; Zhang et al., 2023), we use the rank-based Spearman correlation coefficient as the evaluation metric.

4.1.2. Training Details

We conduct all the experiments on 2 Nvidia RTX A6000 GPUs with PyTorch 2.0.0. The parameter settings of our framework are as follows:

In the backbone model training step, we initialize sentence encoders from SimCSE supervised checkpoints of `BERT-large`, and `RoBERTa-large`, while GenSE checkpoint of `T5-Base`. And we use `RoBERTa-large` for the NLI discriminator.

In the LLM fine-tuning step, we employ WizardLM 7B (Xu et al., 2023) as the small-parameter LLM, and use LoRA (Hu et al., 2021) for efficient finetune, with parameters r = 16 and α = 32. The KL penalty coefficient was configured at 0.2, and the learning rate was set to 1.41×10^{-5}. The weights for the Reward Model, w_1 and w_2, are set to 0.5.

In the sentence encoder training step, we adopt the vLLM (Kwon et al., 2023) for fast sample generation. For each round, we train our sentence encoder for 2 epochs with temperature $\tau = 0.05$.

4.1.3. Baselines

We compare our method to several state-of-the-art approaches based on two existing supervised sentence representation methods, namely SimCSE (Gao et al., 2021b) and GenSE (Chen et al., 2022b). The distinction among these methods lies in the usage of different datasets for training the sentence encoders.

SimCSE uses the NLI dataset as the training data. Based on SimCSE, SimCSE* additionally uses 270K synthesized data generated from WizardLM as the training data; SynCSE (Zhang et al., 2023) only uses 270K synthesized data generated from ChatGPT; while our method additionally uses $3 \times 20K$ synthesized data generated from the optimized WizardLM.

GenSE uses 61M synthetic sentences from C4 and English partitions (Raffel et al., 2020) as the training data. Based on GenSE, GenSE* additionally uses 270K synthesized data generated from WizardLM as the training data; GenSE+ additionally uses 4M QA pairs; while our method additionally uses $3 \times 20K$ synthesized data generated from the optimized WizardLM. PromCSE (Jiang et al., 2022), current SOTA method, is a prompt-based contrastive Learning for sentence embeddings framework.

4.1.4. Research Questions

To evaluate the performance of our method, we design experiments to answer the following research questions:

RQ1: How effective is our method compared to existing baselines?

RQ2: How effective is our adaptive iterative framework?

RQ3: How effective is our dual reward model in the reinforcement tuning framework, especially for generation accuracy and difficulty?

In RQ1, we investigate whether our model outperforms the baselines of sentence representation. In RQ2, we further investigate whether our adaptive iterative framework can continuously improve the sentence encoder. In RQ3, we analyze whether our dual reward model in the reinforcement tuning framework can continuously improve the quality of synthetic data.

4.2. RQ1: Performance Comparison

We perform a performance comparison between our method and several state-of-the-art methods. To be specific, we use `BERT-large`, `RoBERTa-large`, and `T5-base` as pre-trained models, and Table 2 displays the performance on seven STS tasks. The detailed analysis is as follows.

Using `BERT-large` as the base model, our method outperforms all other methods on 4 out of 7 STS datasets and achieves the highest average Spearman correlation across all datasets. Compared to SimCSE, our method shows significant improvements of 1.44-3.22 points on multiple datasets. With `RoBERTa-large`, our method achieves the highest scores on STS12 and STS15,

Methods	Extra Data	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	AVG.
BERT-large									
SimCSE	-	75.78	86.33	80.44	86.06	80.86	84.87	**81.14**	82.21
SimCSE*	270K Raw-LLM	77.50	87.11	81.44	87.09	82.96	85.45	80.74	83.18
PromCSE	-	78.43	87.31	82.09	**87.85**	83.16	**85.62**	80.74	83.60
SynCSE	270K ChatGPT	78.30	87.26	81.27	86.87	82.88	85.44	80.73	83.25
Ours	3*20K RL-LLM	**79.00**	**87.85**	**82.25**	87.42	**83.51**	85.35	80.17	**83.65**
RoBERTa-large									
SimCSE	-	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
SimCSE*	270K Raw-LLM	79.98	87.57	82.80	86.67	84.64	86.03	81.58	84.18
PromCSE	-	79.14	**88.64**	**83.73**	87.33	84.57	**87.84**	82.07	**84.76**
SynCSE	270K ChatGPT	77.13	87.61	82.82	87.67	**85.66**	87.22	**82.45**	84.37
Ours	3*20K RL-LLM	**80.38**	88.63	83.61	**87.70**	85.05	86.45	80.78	84.66
T5-base									
GenSE	-	80.72	87.43	83.96	88.63	85.19	87.65	79.87	84.78
GenSE*	270K Raw-LLM	80.84	87.54	84.23	88.72	85.31	87.72	79.63	84.86
GenSE+	4M QA (Real)	80.65	**88.18**	**84.69**	**89.03**	**85.82**	**87.88**	80.10	**85.19**
Ours	3*20K RL-LLM	**81.21**	87.93	84.41	88.83	85.36	87.87	**80.21**	85.12

Table 2: Results on seven STS datasets. (**Bold**: the best. Underlined: the second best.)

Base Model	Methods	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	AVG.
BERT-large	Round 0	75.78	86.33	80.44	86.06	80.86	84.87	**81.14**	82.21
	Round 1	78.36	87.26	81.66	87.22	83.15	85.51	80.85	83.43
	Round 2	78.66	87.64	82.08	87.39	83.37	**85.52**	80.47	83.59
	Round 3	**79.00**	**87.85**	**82.25**	**87.42**	**83.51**	85.35	80.17	**83.65**
RoBERTa-large	Round 0	77.46	87.27	82.36	86.66	83.93	86.70	**81.95**	83.76
	Round 1	78.98	88.43	83.54	87.61	85.00	**86.81**	81.25	84.52
	Round 2	79.42	**88.64**	**83.62**	87.69	**85.27**	86.64	80.95	84.60
	Round 3	**80.38**	88.63	83.61	**87.70**	85.05	86.45	80.78	**84.66**
GenSE-T5-Base	Round 0	80.72	87.43	83.96	88.63	85.19	87.65	79.87	84.78
	Round 1	81.07	87.80	84.33	88.75	85.34	87.75	79.81	84.98
	Round 2	**81.42**	**88.18**	**84.53**	88.67	**85.44**	87.53	79.70	85.07
	Round 3	81.21	87.93	84.41	**88.83**	85.36	**87.87**	**80.21**	**85.12**

Table 3: Results of each round on seven STS datasets.

and the highest average score among fine-tuning models. Notably, it outperforms SimCSE consistently on STS12-16, showing gaps of 1.12-2.92 points. This again verifies the advantage of our method over raw LLM data. Compared to ChatGPT data, our method achieves better overall performance, despite being weaker on STS-B and SICK-R. In comparison, PromCSE enjoys better performance with the prompt-based method (namely P-Tuning v2) on STS-B and SICK-R datasets. Such method optimizes the model by adding additional parameters, resulting in a sentence representation that is different from the checkpoint in our dual reward model. Thus, our sentence encoder cannot integrate prompt-based method. For the smaller `T5-base` model, our method surpasses the GenSE baseline on all datasets, with improvements of 0.34-0.79 points. Compared to GenSE+ which uses 400M additional QA data, our method achieves comparable results using much less synthetic data. This shows the data efficiency of our approach.

In all, our method achieves strong overall results with far fewer data, yet achieves results comparable to or even surpassing those obtained by directly using large-parameter LLMs.

4.3. RQ2: Performance Comparison of Adaptive Iterative Framework

To demonstrate the benefits of our adaptive iterative framework, we provide an overview of the results achieved at each round. Table 3 presents the performance of three pre-trained models on seven STS tasks at each round of training.

Our results show an adaptive enhancement in the model's performance with each successive round. As expected, a stronger baseline gains fewer performance upraise. Specifically, for `BERT-large`, at round 1, round 2, and round 3, the average Spearman correlations in the seven STS datasets are 1.22%, 1.38%, and 1.44% higher than SimCSE, respectively. For RoBERTa-large, the gains are 0.76%, 0.84%, and 0.9% higher correlations versus SimCSE for the three rounds. Additionally, based on `T5-Base`, a stronger baseline, gains plateau at 0.34% higher correlation. In those mod-

Round	Pos. Acc.	Neg. Acc.	Avg. Acc.
Round 0	85.96%	78.72%	82.34%
Round 1	92.89%	95.67%	94.27%
Round 2	**93.54%**	96.52%	95.03%
Round 3	93.51%	**96.68%**	**95.09**%

Table 4: Result of synthetic sample accuracy, based on RoBERTa-large sentence encoder.

Syn Data	Wizard	SynCSE	GenSE	Ours
x, x^+ **cos**$\downarrow$	0.94	0.91	0.85	**0.84**
x, x^- **cos**$\uparrow$	0.57	0.60	0.51	**0.83**

Table 5: Result of average synthetic sample difficulty, ours from round 3, calculated by supervised SimCSE RoBERTa-large, Wizard stands for WizardLM.

els, round 3 always contributes the smallest gains, thus, we choose to stop at this round. From the perspective of the variation between tasks, our results show a consistent improvement in the STS tasks after multiple rounds of training. This maximizes the utility of the data generated. We observed a consistent decline in the SICK-R task, up to 1.0. We believe that this discrepancy may be due to biases in the distribution of our synthesized data. These issues are also reflected in the performance of GenSE+ (Chen et al., 2022b) models. In all, these experimental results validate the effectiveness of our adaptive iterative framework.

4.4. RQ3: Synthetic Data Quality

4.4.1. Ablation Study 1: Accuracy of Synthetic Data

Here we examine the effectiveness of NLI discriminator in our dual reward model. Table 4 shows the accuracy of the synthetic positive and negative samples improves over multiple rounds of adaptive training. After round 1, the positive accuracy increases to 92.89% from 85.96% for the raw LLM. The negative accuracy rises to 95.67% from 78.72%. The overall average accuracy improves by 12.33% to 94.27%. In round 2, the positive accuracy reaches 93.54% and the overall average hits 95.03%. By round 3, the negative accuracy peaks at 96.68%, and the overall average reaches 95.09%. The steady accuracy improvements verify that with the dual reward model, our method not only generates more challenging samples but also significantly improves content accuracy. With the dual reward model incorporated, the steady improvement in accuracy confirms that our method not only generates more challenging samples but also significantly enhances content accuracy.

4.4.2. Ablation Study 2: Difficulty of Synthetic Data

Here we examine the effectiveness of the difficulty evaluator in our dual reward model. Shown in Table 5, we compare the ability of different LLMs to create challenging samples for semantic textual similarity. Lower cosine similarity indicates greater difficulty for positives, and vice versa. Our model achieves a lower average positive cosine similarity score of 0.84, indicating it creates the most challenging positives. Most importantly, our model obtains the highest average negative cosine similarity score of 0.83, meaning it generates negatives of the highest difficulty compared to using frozen LLMs and previous synthetic methods like GenSE. In summary, our proposed model produces harder samples which better enhances the capability of semantic similarity modeling.

4.4.3. Case Study

Here is the overall analysis of the case study conducted on the hard positive and negative sentences generated in different rounds, as shown in Table 6. Several trends emerge. First, the generated samples aim to preserve sentence structure and vocabulary as much as possible, while making minimal edits to convey entailment or contradiction. Lexically, negative samples primarily involve antonym substitutions like changing “sitting" to “standing" and color adjective changes like changing “girl" to “boy". Such positive samples provide additional details like mentioning the slope of the mountain.

Syntactically, negative samples have minimal variations to appear highly similar and difficult to distinguish from the anchor. Positive samples exhibit greater syntactic changes like modifying tense and switching subject and object positions. For the first case, the positive sample of round 1 is a simple change in word order. In contrast, the positive samples of round 2 and 3 modifies the syntax significantly, by converting to passive voice and separating the subject (“a girl") from the main clause. In essence, the adaptive training pushes the LLM beyond simple word order changes to produce varied lexical, syntactic, and semantic paraphrases of the anchor sentence.

5. Related Work

5.1. Contrastive Learning for Sentence Representation

Existing sentence representation methods primarily use the contrastive learning (CL) paradigm to learn effective sentence representations. These methods can be divided into unsupervised and supervised approaches. Unsupervised methods

Round	Positive Sample	Negative Sample
	Anchor: *A girl is sitting on the side of a mountain.*	
Round 1	Aside the mountain, a girl is sitting.	A girl is standing on the side of a mountain.
Round 2	She sat on the side of the mountain, a girl.	A boy is sitting on the side of a mountain.
Round 3	The slope of the mountain is where a girl is seated.	Boy is sitting on the side of a mountain.
	Anchor: *A man is kneeling down and using a paintbrush.*	
Round 1	Painting with a brush, a man knelt down.	A man is standing up and using a paintbrush.
Round 2	He is painting with a brush while standing on his knees.	A paintbrush is kneeling down and using a woman.
Round 3	A man is on bended knee, daubing with his brush.	A woman is kneeling down and using a paintbrush.

Table 6: Comparison of different data synthesis results at different rounds for RoBERT-large

use various data augmentation strategies to generate contrastive data. For example, SimCSE (Gao et al., 2021b) uses dropout to generate positive pairs while taking other sentences as negatives. ESimCSE (Wu et al., 2022b) employs word repetition positives and retrieved negatives data augmentation strategies. However, by using human-annotated natural language inference (NLI) data, the supervised models significantly outperform the unsupervised models. GenSE (Chen et al., 2022b), stands as a semi-supervised sentence representation learning approach that generates and discriminates a substantial dataset from a single T5 model (Raffel et al., 2020). SynCSE (Zhang et al., 2023), synthesizes data from ChatGPT (Ouyang et al., 2022) for training sentence embeddings, demonstrating the potential of utilizing LLM-generated datasets for this task.

5.2. Data Augmentation with LLMs

Data augmentation is a popular technique in NLP that involves generating new text through the application of various transformations. Recently, large language models (LLMs), such as ChatGPT and GPT4 (Ouyang et al., 2022; OpenAI, 2023), have become available. This opened up new possibilities for automatically generating high-quality synthetic contrastive data. Existing methods rely heavily on large-parameter LLMs for data augmentation. (Dai et al., 2023) propose a text data augmentation approach based on ChatGPT to improve language comprehension abilities. (Zhang et al., 2023) synthesizes data from ChatGPT for training sentence embeddings. These methods require crafting multiple prompts to generate proper samples, which is costly at scale. Using these models leads to substantial API expenses or high costs for local deployment (Hsieh et al., 2023). Even when using medium-parameter LLMs, the cost remains high. However, using small-parameter LLMs (Touvron et al., 2023; Xu et al., 2023) for data augmentation has not been thoroughly explored. Still, the quality of data generated by small-parameter LLMs is considerably poorer than larger ones. Therefore, we propose a more cost-effective strategy to optimize small-parameter LLMs to generate synthetic data.

5.3. Reinforcement Learning in LLMs

Recently, extending pre-trained language models by increasing the number of parameters, and training data (Kaplan et al., 2020) can make LLM powerful in various language tasks (Brown et al., 2020). Moreover, recent investigations have further identified the potential of LLMs through supervised fine-tuning (SFT) and reinforcement learning based on human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023). Our approach is inspired by the RLHF method. RLHF was originally developed for training simple robots in simulated environments and games (Christiano et al., 2017). It also is applied to fine-tuning language models to summarize text (Ziegler et al., 2019). With RLHF, language models can be better aligned with human preferences, i.e., better follow human instructions. Learning improved language models from human feedback through reinforcement learning techniques has been explored in (Ouyang et al., 2022; Korbak et al., 2023). Most of the existing studies employ the PPO algorithm to fine-tune LLMs (Schulman et al., 2017). Here, we introduce a novel reinforcement learning approach for fine-tuning small-parameter LLMs, enabling the generation of high-quality contrastive sentence data without human feedback.

6. Conclusion

In this paper, we proposed a cost-effective strategy to utilize small-parameter LLMs to generate synthetic data and enhance sentence representation models. Specifically, we propose a novel adaptive reinforcement tuning (ART) framework to optimize small-parameter LLMs in a reinforcement learning manner to adaptively generate hard contrastive samples with multiple iterations. These samples

are then used to enhance the performance of existing sentence representation models. Experiments conducted on seven semantic text similarity tasks demonstrate that the sentence representation models trained using the synthetic data generated by our proposed method achieve state-of-the-art performance. We also conducted ablation studies to showcase the critical roles played by the reinforcement learning approach and the adaptive iterative framework within our proposed framework.

7. Acknowledgements

We are very grateful to the anonymous reviewers for their hard work and valuable comments. This work is supported by the Fundamental Research Funds for the Central Universities 2232023D-19 and the NSF of Shanghai under grant number 22ZR1402000. Ming Du is the corresponding author.

8. Bibliographical References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. 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)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval@ COLING*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. 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)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, pages 1–16.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Zhengda Bian, Hongxin Liu, Boxiang Wang, Haichen Huang, Yongbin Li, Chuanrui Wang, Fan Cui, and Yang You. 2021. Colossal-ai: A unified deep learning system for large-scale parallel training. *arXiv preprint arXiv:2110.14883*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual

and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022a. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yiming Chen, Yan Zhang, Bin Wang, Zuozhu Liu, and Haizhou Li. 2022b. Generate, discriminate and contrast: A semi-supervised sentence representation learning framework. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. 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*, pages 4207–4218.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2022. Learning general halfspaces with general massart noise under the gaussian distribution. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 874–885.

Kawin Ethayarajh. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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 55–65.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

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

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Cheng-Yu Hsieh, Chun-Liang Li, and Chih-kuan Yeh. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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*.

Sverker Janson, Evangelina Gogoulou, Erik Ylipää, Amaru Cuba Gyllensten, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations, 2021*.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of human preferences in dialog.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3021–3035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

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

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 1–10.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Renhao Li, Lei Duan, Guicai Xie, Shan Xiao, and Weipeng Jiang. 2022. Adcse: An adversarial method for contrastive learning of sentence embeddings. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III*, pages 165–180.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. 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)*, pages 216–223.

Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. 2021. Self-supervised contrastive learning with adversarial perturbations for robust pretrained language models. *arXiv preprint arXiv:2107.07610*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith 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, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

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.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Marco Damonte, Isabel Groves, and Amir Saffari. 2022. Clasp: Few-shot cross-lingual data augmentation for semantic parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 444–462.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for

face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Jie Shuai, Kun Zhang, Le Wu, Peijie Sun, Richang Hong, Meng Wang, and Yong Li. 2022. A review-aware graph contrastive learning framework for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1283–1293.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Afrina Tabassum, Muntasir Wahed, Hoda Eldardiry, and Ismini Lourentzou. 2022. Hard negative sampling strategies for contrastive representation learning. *arXiv preprint arXiv:2206.01197*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bin Wang, C-c Kuo, and Haizhou Li. 2022a. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022b. Sncse: contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wei Wang, Liangzhu Ge, Jingqiao Zhang, and Cheng Yang. 2022c. Improving contrastive learning of sentence embeddings with case-augmented positives and retrieved negatives. *arXiv preprint arXiv:2206.02457*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Yipeng Su, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022a.

Smoothed contrastive learning for unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4902–4906.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3898–3907.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. 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 (Volume 1: Long Papers)*, pages 5065–5075.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew Arnold. 2022a. Virtual augmentation supported contrastive learning of sentence representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 864–876.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2023. Contrastive learning of sentence embeddings from scratch. *arXiv preprint arXiv:2305.15077*.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. 2022b. Mcse: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931*.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J Passonneau. 2022c. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022d. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11730–11738.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022e. 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 (Volume 1: Long Papers)*, pages 4892–4903.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.