Empowering Sentence Encoders with Prompting and Label Retrieval for Zero-shot Text Classification

Anonymous ACL submission

Abstract

With contrastive pre-training, sentence encoders are generally optimized to locate semantically similar samples closer to each other in their embedding spaces. In this study, we empirically investigate the adaptability of these embedding spaces for zero-shot text classification, given that semantically distinct samples are already well-separated. Remarkably, the simple approach utilizing a sentence encoder to assign labels based on the highest similarity between the prompt embedding and the input text embedding showcases impressive zero-shot performance. Nonetheless, the approach is hindered by inadequately informative label prompts in their initial form. Therefore, we harness the same sentence encoder to extract semantically similar label prompts from external corpora and employ them as supplementary pseudo-label prompts. Altogether, the approach demonstrates stronger performance than state-of-the-art baselines on various closed-set classification and multiple-choice QA datasets under zero-shot settings. We analyze that the retrieval component plays a pivotal role in success and its results are robustly attained regardless of verbalizer variations.

1 Introduction

Sentence encoders have been widely applied to a comprehensive range of natural language processing tasks, including classification, semantic retrieval, and semantic similarity tasks (Reimers and Gurevych, 2019; Du et al., 2021; Gao et al., 2021b; Ni et al., 2022). They are usually pre-trained with a contrastive objective on datasets that focus on sentence semantics (*e.g.* NLI), so semantically similar texts are located close to each other in their embedding spaces. We note that embedding spaces with such traits could be particularly friendly to classification tasks under limited supervision, as semantically distinct samples are well-separated in advance of any refinement made during downstream training. Recent work has demonstrated the competitiveness of rich text embeddings from sentence encoders for few-shot classification (Tunstall et al., 2022), being on par or outperforming much larger prompt-based generative language models on the RAFT benchmark (Alex et al., 2021). However, it essentially involves training a logistic classification head, which introduces additional parameters to be tuned and is thus inapplicable to zero-shot inference.

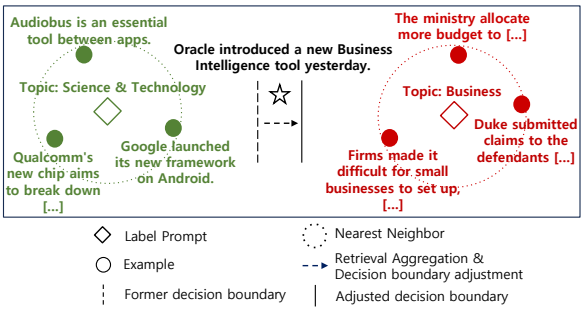


Figure 1: In the visualization of embedding space, diamond-shaped label prompts ("Topic: Science & Technology" and "Topic: Business") serve as anchors for the classification. By utilizing retrieved samples from external corpora in proximity to each label prompt, the decision boundary can be adjusted accordingly.

In this work, we continue to explore the power of sentence encoders, but this time to solve zero-shot classification by combining with prompting (Brown et al., 2020) and dense retrieval techniques. We transform closed or open-set classification tasks into finding the label texts (*i.e.* prompts) with maximal textual similarity with the sentences under inference in a sentence encoder's embedding space. Meanwhile, we note that label prompts in their original format are potentially ambiguous or poorly descriptive; for example, *"Topic: Science & Technology."* in the AGNews dataset is an overly compressive and abstract label, being less adequate to embrace a large range of belonging texts. We handle this issue by using additional support from multiple pseudo-label prompts retrieved from an external corpus, as illustrated in Figure 1.

Our approach distinguishes itself from previous works in zero-shot classification by comparing the similarities of textual embeddings between label prompts and input samples. Previous works using prompting with generative language models (Brown et al., 2020; Gao et al., 2021a) have exhibited high variance across different verbalizers (i.e., text expressions for labels), even being close to chance-level performance (Perez et al., 2021; Lu et al., 2022). In contrast to generative models that intertwine label prompts and query texts through concatenation at the input level, our approach isolates the two during encoding. Through empirical evidence, we demonstrate that such post-hoc interactions can withstand variations in label prompts, and this robustness is further enhanced when we expand the support prompt set via retrieval.

We conduct extensive experiments on six closed-set classification datasets and six open-set multiple-choice datasets. The best setup which uses a sentence encoder *Sentence-T5* (Ni et al., 2022) with retrieval augmented label prompts achieves comparable or stronger performance compared to state-of-the-art baselines for both tasks. Furthermore, the approach is consistently strong across verbalizer variations and scales to different model sizes. Our contributions are three-fold:

- We empirically investigate how sentence encoders perform in a zero-shot classification setup, which surpasses strong baselines
- We suggest a retrieval module to address the problem of vague label prompts, a crucial aspect in a zero-shot setup
- We analyze the approach's utility in terms of retrieving pseudo-label prompts and resilience to verbalizer variations

2 Background

We explain the concept of sentence encoders and prompting that we use in this work.

2.1 Sentence Encoders

Given a sentence $X_i \in \mathcal{X}$, a sentence encoder E encodes the sentence into a fixed size embedding vector $\mathbf{h}_i = E(X_i) \in \mathbb{R}^d$, where $\mathcal{X}$ is the set of all natural language texts and d is the preset embedding dimension. Sentence encoders are commonly trained with the contrastive learning objective (Chen et al., 2020) with in-batch negatives (Chen et al., 2017; Henderson et al., 2017). The loss function pulls the positive pair representation closer to the input representation while pushing away the negatives in embedding space:

$$\ell = -\sum_{i=1}^{N} \log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)}}{\sum_{j=1, j\neq i}^{N} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j)}} \quad (1)$$

where $\mathrm{sim}(\cdot)$ is the similarity function, $\mathbf{h}_i^+$ denotes the positive pair for $\mathbf{h}_i$, and $\mathbf{h}_j$ refers to all other instances except $\mathbf{h}_i$ in the batch of size N.

2.2 Prompting

Given a text input $X_i \in \mathcal{X}$ and a set of labels $\mathcal{Y} = \{y_1, \ldots, y_m\}$, a predefined verbalizer[1] $v : \mathcal{Y} \rightarrow \mathcal{X}$ generates a label prompt in natural language for each label index. Generally, language models compute the distribution of label prompts given the input, $P_{\mathrm{LM}}\left(v\left(y_m\right) \mid X_i\right)$ (Brown et al., 2020). Other lines of works utilize channel modeling $P_{\mathrm{LM}}\left(X_i \mid v\left(y_m\right)\right)$ (Min et al., 2022a,b) or masked language modeling objective (Gao et al., 2021a).

3 Method

We explore solving zero-shot classification tasks using text representations from sentence encoders. Figure 2 provides an overview of our approach. The gist is to exploit the representational similarity between the input text and retrieval-augmented label prompts.

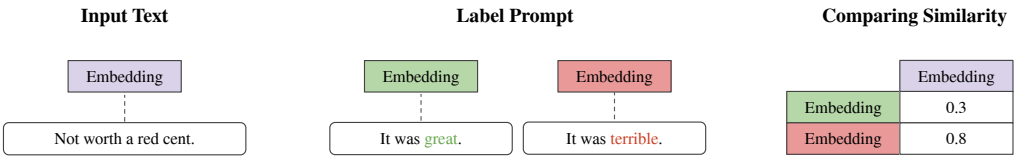
3.1 Representational Similarity with Sentence Encoders

Taking the example in Figure 2 (a), for a binary sentiment classification task with $\mathcal{Y} = \{y+, y-\}$ where $y+$ stands for "positive" and $y-$ stands for "negative", we first use a verbalizer to change each label into prompts: say, *"It was great."*$(v(y+))$ and *"It was terrible."*$(v(y-))$. Afterward, the label prompts are transformed into encoded vectors that act as prototypes for representing each class cluster: $\mathbf{z}_m = E(v(y_m))$. Then, we use the following scoring function for each candidate label, defined as its similarity with the text under inference X_i *"Not worth a red cent."*:

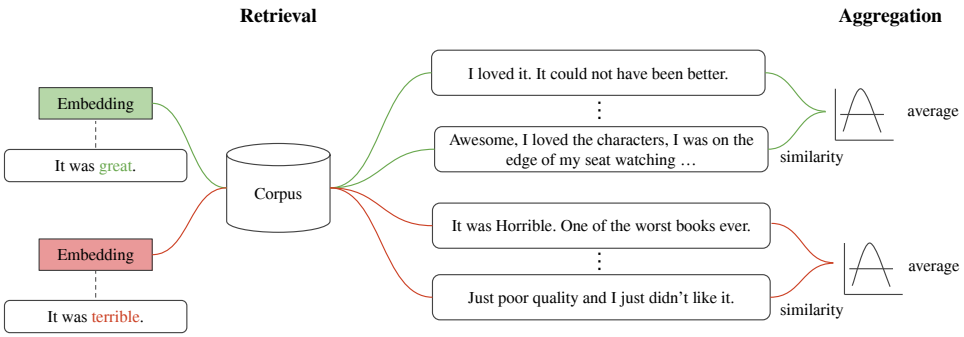
$$\underset{y_m \in \mathcal{Y}}{\mathrm{argmax}} \ \mathrm{P_{TS}}(y_m|X_i)$$
$$\mathrm{P_{TS}}(y_m|X_i) \propto \mathrm{sim}(\mathbf{h}_i, \mathbf{z}_m) \quad (2)$$

where we use cosine similarity in a sentence encoder's embedding space for $\mathrm{sim}(\cdot)$.

[1] Consists of templates and label names.



(a) Comparing Textual Similarity with Label Prompting



(b) Retrieval Augmented Label Prompts

Figure 2: An overview of our approach to a text classification task (or input-label matching). Among a set of label candidates, we find the label whose prompt has the highest representational similarity with the input text in the embedding space of a sentence encoder (Figure (a)). The similarity scores are computed across multiple retrievals augmented label prompts which are collected with original label prompts as queries (Figure (b)).

Note that using Eq. 2 at inference time aligns the downstream task objective with the contrastive pre-training objective (Eq. 1), which may benefit sentence encoders in a limited label regime. Plus, separately encoding label prompts from the test input makes the results more robust under verbalizer variations, which we show in Section 6.1.

3.2 Retrieval Augmented Label Prompts

It is crucial for the label prompts to reliably express the necessary class information if each of them is to act as the single similarity anchor for its class. Unfortunately, this precondition often does not hold as many label names are being compressive rather than fully descriptive; *e.g. "Topic: World"* in the AGNews dataset.

To make amends for the potential issue of fragmented semantics in the labels in their original format, we augment the label prompts with semantically similar sentences retrieved from external knowledge sources. Texts collected from the wild are expectedly more descriptive. Therefore, by serving as alternative formats or augmentations to the original label prompts, they may add useful class-related information, such as synonyms or related expressions. We would like to highlight that the retrieval augmentation technique employed in our approach is specifically designed to address the problem of inadequate descriptive class information, setting it apart from other existing methods.

We retrieve top-K sentences from an external corpus according to the representational similarity with the given label prompt. Then, we compute the aggregated score for the retrieved *set of sentences* as

$$\mathrm{P}_{\mathrm{ES}}(y_m|X_i) \propto \frac{1}{K}\sum_{k=1}^{K}\mathrm{sim}(\mathbf{h}_i, \mathrm{topk}(\mathbf{z}_m)^k) \quad (3)$$

where K refers to preset retrieval size, topk is an operator which returns top-k nearest embeddings from encoded external sources, and ^{k} denotes the k-th element from the retrieval results. The sum of Eq. 2 and Eq. 3 now becomes our new inference scoring function.

$$\underset{y_m \in \mathcal{Y}}{\mathrm{argmax}} \ \mathrm{P}_{\mathrm{TS}}(y_m|X_i) + \mathrm{P}_{\mathrm{ES}}(y_m|X_i)$$

For further augmentation during retrieval, we experiment with using synonyms of the original label names (*e.g. "good", "remarkable"* for *"great"*) as in Shi et al. (2022). Given a label prompt, we first generate a size-N list of synonymous label prompts by replacing the label name with its synonyms. Then, we do the retrieval for N times, each time with a formerly generated synonymous prompt as the query and K/N retrieved sentences.

4 Experimental Settings

We experiment with two types of tasks: closed-set classification and multiple choice QA. Both involve choosing the corresponding label given a finite label set for each test sample; what tells them apart is whether all test instances share the same label set (closed-set classification) or each test instance is presented with a different each label set (multiple choice QA).

4.1 Closed-set Classification Datasets

Sentiment Analysis SST-2 (Socher et al., 2013), MR (Pang and Lee, 2004), CR (Hu and Liu, 2004), and Rotten Tomatoes (RT) (Socher et al., 2013).

Topic Classification AGNews (news domain) and Yahoo Answers (Yahoo; web domain) (Zhang et al., 2015).

Details We experiment on the modified test set used in Shi et al. (2022); Min et al. (2022c).[2] We report the scores averaged over four templates (Gao et al., 2021a; Min et al., 2022a). See Table 5 for the list of verbalizers used.

4.2 Multiple Choice QA Datasets

We evaluate on RACE-M (R-M), RACE-H (R-H) (Lai et al., 2017), ARC-E, ARC-C (Clark et al., 2018), Open Book Question Answering (OBQA) (Mihaylov et al., 2018), and CommonsenseQA (CoQA) (Talmor et al., 2019). For CoQA, we report the validation results as the official test set is not publicly available. We use a single template from Holtzman et al. (2021). See Table 6 for details on templates and verbalizers.

4.3 Baseline Models

Standalone decoder language models **(LM)** select the label whose word sequence has with the highest probability. We use *gpt-2-large* (Radford et al., 2019) and variants of *gpt-3* (Brown et al., 2020). **PMI** (Holtzman et al., 2021) calibrates the decoder models with a domain-conditioned premise. We also include an encoder-based cloze-style zero-shot classification model with RoBERTa (Liu et al., 2019) as the encoder. For retrieval-augmented baselines, we use **kNN-LM** (Khandelwal et al., 2019) and **kNN-prompt** (Shi et al., 2022), which adjust the output token probabilities with external resources. We also include **NPM** (Min et al., 2022c) as our baselines which uses a non-parametric masked language model for phrase-level retrieval. Another approach, proposed by Yin et al. (2019), treats label candidates as hypotheses and selects the label with the highest entailment logit score (**Entailment**). We implement the method by fine-tuning pre-trained *roberta-large* (Liu et al., 2019) on NLI (SNLI + MNLI) datasets (Bowman et al., 2015; Williams et al., 2018).[3]

Although the baseline approaches are not optimized with sentence encoding, they exhibit impressive performance by achieving state-of-the-art results in zero-shot classification tasks, as demonstrated in Shi et al. (2022); Min et al. (2022c). Furthermore, it is worth noting that the zero-shot performance of **RoBERTa** (Gao et al., 2021a) is outstanding, surpassing that of large-scale language models (Gao et al., 2021a; Min et al., 2022c), even though the original work primarily focuses on showcasing the few-shot performance. Additionally, due to the non-trivial nature of adapting **kNN**-based methods for multiple-choice QA tasks, we decided not to include these two approaches. One such example is **kNN-Prompt**, which incorporates Glove embeddings and ConceptNet-based synonyms to enhance verbal descriptions for each class. However, integrating such techniques into solving multiple-choice QA tasks is not a straightforward process.

Moreover, in the context of using sentence encoders, our study includes conducting experiments with the application of **SimPTC** (Fei et al., 2022). These experiments necessitate the use of unlabeled train and test dataset. To ensure a fair comparison, we employ a selection of instances from external corpus. We replicate the SimPTC approach in two distinct manners: 1) by randomly choosing a subset from the external corpus and training the model on these samples, and 2) by training on retrieved augmented label prompts (RaLP).

4.4 Implementation Details

Sentence Encoder For the main experiments, We use two off-the-shelf sentence encoder models: *sup-simcse-roberta-large* (SimCSE) (Gao et al., 2021b) trained on NLI, and *sentence-t5-large* (ST5) (Ni et al., 2022) trained on CommunityQA and NLI. Among the ST5 variations, we use an encoder-only mean version. We also experiment

[2] They made a random subset of 3,000 samples for datasets with larger original test sets.

[3] We use the script provided by sentence-transformers (Reimers and Gurevych, 2019).

Method	# Params	SST-2	MR	CR	RT	Yahoo	AGNews	Avg
Baselines (Language Modeling)								
GPT-2†	774M	55.3	54.6	66.2	53.0	49.7	67.4	57.7
+ PMI† (Holtzman et al., 2021)	774M	76.5	74.6	82.8	74.1	48.8	65.1	70.3
GPT-3†	175B	63.6	57.4	53.8	57.0	53.1	75.4	60.1
+ PMI† (Holtzman et al., 2021)	175B	71.4	76.3	70.0	75.5	54.7	74.7	70.4
RoBERTa (Gao et al., 2021a)	355M	83.2	80.8	79.6	82.6	44.9	68.3	73.2
Baselines (Retrieval-Augmented LM)								
GPT-2 + kNN† (Khandelwal et al., 2019)	774M	55.4	56.4	67.2	54.5	49.5	67.0	58.3
GPT-2 + kNN-Prompt† (Shi et al., 2022)	774M	84.2	78.2	84.3	80.6	51.0	**78.8**	76.2
NPM†	355M	87.2	83.7	81.2	**86.0**	53.9	74.5	77.8
Baselines (NLI format)								
Entailment (Yin et al., 2019)	355M	83.7	79.6	83.8	78.2	46.0	75.1	74.4
Baseline (Sentence Encoder)								
$\text{SimPTC}^{\text{random}}{}_{\text{ST5}}$	335M	76.7	75.7	75.0	69.4	32.3	60.2	64.9
$\text{SimPTC}^{\text{RaLP}}{}_{\text{ST5}}$	335M	85.2	80.9	87.3	78.3	52.9	68.9	75.6
Ours								
$\text{RaLP}_{\text{SimCSE}}$	355M	82.3	78.0	87.1	76.7	57.0	72.6	75.6
RaLP_{ST5}	335M	**87.8**	**81.7**	**87.4**	82.4	**57.4**	76.6	**78.9**
$\text{RaLP}_{\text{ST5-XL}}$	1.24B	88.6	82.8	86.1	83.3	56.2	75.1	78.7
$\text{RaLP}_{\text{ST5-XXL}}$	4.8B	90.5	84.7	87.7	85.1	54.9	75.4	79.8

Table 1: Results for closed-set classification tasks under the zero-shot setting. The **Boldface** indicates the best performances in each column except the models with over 1B parameters. The Underline shows the best scores in each column regardless of model size. Values with † are taken from Min et al. (2022c).

with larger versions of ST5 (*sentence-t5-xl* (ST5-XL) and *sentence-t5-xxl* (ST5-XXL)) to see the effect of model size.

External Corpus We collect the external corpus used in Shi et al. (2022) which consists of Wiki103, IMDB, subsets of CC-News, and Amazon Review.

Retrieval We use the same sentence encoders as Section 4.4 for building the embedding index, and use FAISS (Johnson et al., 2019) for the top-k computation and retrieval. For each original label prompt, we use 5 synonymous prompts and retrieve 5 sentences per query (so that K=25) in closed-set classification. See Table 8 and Table 9 for a list of retrieved examples. For multiple choice tasks, we use 25 retrieved sentences for a single query. We do not use synonymous label prompts as making small changes in the wordings may deviate from the originally intended semantics, thus making the option less of a viable answer. We observe no further gain beyond 25 retrieved samples.

5 Results

Our method shows solid performance on both task types under zero-shot setting.

5.1 Closed-set Classification

Table 1 shows the results for closed-set classification tasks.

Baselines On average, NPM shows the strongest results among all baselines. This indicates the importance of extracting and exploiting relevant information from external knowledge sources in zero-shot inference. Among the baselines that do not make use of external resources, RoBERTa outperforms the standalone decoder models with or without PMI calibration, presumably due to utilizing bi-directional information with an encoder.

Among the approaches using the sentence encoder, our observations reveal a noteworthy decline in performance when SimPTC is applied to unlabeled examples. This highlights that SimPTC's effectiveness relies on access to an in-domain labeled dataset, a setup that proves difficult and impractical. However, we found that employing SimPTC with RaLP can effectively mitigate the performance decline.

RaLP Our method shows strong performance when coupled with a well-built sentence encoder. With the ST5 backbones, we outperform all the baselines despite fewer parameters.[4] Changing the sentence encoder to SimCSE results in a lower performance, but still comparable to RoBERTa.

RaLP serves as a cost-effective method of leveraging retrieval from external sources. Although KNN-Prompt (Shi et al., 2022) and NPM (Min

[4]We attribute ST5's strong zero-shot classification abilities to its architecture rather than the pretraining dataset composition. See appendix A for details.

Sentence Encoder	ST5		SimCSE	
RaLP	✗	✓	✗	✓
SST-2	87.7	**87.8**	80.8	**82.3**
MR	**81.7**	**81.7**	76.5	**78.0**
CR	87.3	**87.4**	**87.5**	87.1
RT	81.9	**82.4**	75.7	**76.7**
Yahoo	55.3	**57.4**	53.4	**57.0**
AGNews	70.4	**76.6**	67.6	**72.6**
Avg	77.4	**78.9**	73.6	**75.6**

Table 2: Ablation study on utilizing RaLP for closed-set classification tasks.

et al., 2022c) depend on the same external corpus as ours, RaLP outperforms KNN-Prompt not only in terms of task accuracy[5], but also in terms of computational efficiency.

Specifically, whereas KNN-Prompt requires token-level encoding of the leftward context and the next token, we employ a much lighter instance-level encoding. Such difference in the encoding strategy allows RaLP to operate on merely 2M items while KNN-Prompt and NPM need to store 284M, and 188M items taking external corpus mentioned in Section 4.4.

Ablation on Retrieval Augmentation We evaluate the effects of using retrieval augmented label prompts. From Table 2, we observe that using supplementary label information results in a solid performance increase. The degree of performance improvement by retrieval indeed varies depending on the nature of each task. These observations suggest that tasks with topic-associated aspects like Yahoo and AGNews show more prominent performance enhancement through retrieval compared to sentiment-related tasks, namely SST-2, MR, CR, and RT. In topic-related tasks where labels (e.g., "world") may not encapsulate all relevant context, retrieval tends to provide a substantial advantage in bridging this information discrepancy. Conversely, in sentiment tasks where labels tend to have direct semantic alignment, the incremental gains from retrieval might be less noticeable.

5.2 Multiple Choice QA

Table 3 shows the results for multiple choice QA tasks.

Baselines In sharp contrast to the results in Section 5.1, the decoder approaches show strong performance with dramatic gains from upscaling. The highest scores come from a giant decoder model (175B GPT-3+PMI), presumably due to massive memorization of specific knowledge.

RaLP Still, our method based on sentence encoders still achieves better performance than the baselines with an equivalent (or larger) number of parameters. RaLP_{ST5} (with 335M parameters) outperforms GPT-2+PMI (which has the best performance with $\leq$ 774M parameters) by 2.5 points in average. Our best model $\text{RaLP}_{\text{ST5-XXL}}$ (with 4.8B parameters) outperforms much larger GPT-3+PMI baselines (with 6.7B and 13B parameters) by 7.9 and 3.3 points in average; even on par with a 175B GPT-3 despite having 36.5x fewer parameters.

6 Analysis

6.1 Verbalizer Sensitivity Test

Existing works based on prompting are vulnerable to verbalizer changes, with worst-case performances often down to chance-level (Lu et al., 2022). We suggest that using sentence encoders may be a remedy, as they are trained to distribute semantically similar samples nearby during contrastive pre-training and thus could be less sensitive to surface form variations. In this section, we empirically verify that our method based on sentence encoders is *reliably* strong.

Settings We measure the variance in performance across a range of paraphrased label templates while keeping the label words intact: say, changing *"It was great."* for sentiment analysis to *"It's a great thing."* or *"That's great."*. For paraphrasing, we leverage templates from existing works (Gao et al., 2021a; Min et al., 2022a) and augmentation techniques (Ma, 2019) including back-translation (Sennrich et al., 2016) and contextual word embedding-based augmentation (Kobayashi, 2018). Among the generated candidates, we manually filter out the ones with semantic distortions (*e.g.* added negations).[6] We use *roberta-large* for the baselines (Yin et al., 2019; Gao et al., 2021a). For fair comparison, we share the same backbone with the baselines and use *sup-simcse-roberta-large* (SimCSE) (Gao et al., 2021b) as our sentence encoder.

Results Figure 3 shows the accuracy distributions over verbalizer variations. As previously re-

[5] RaLP_{ST5} (335M) reaches a higher performance than KNN-Prompt which has twice as more parameters (774M)).

[6] See Table 7 for a full list.

Method	# Params	R-M	R-H	ARC-E	ARC-C	OBQA	CoQA	Avg
Baselines (Language Modeling)								
GPT-2	774M	39.3	31.8	52.7	23.1	19.4	33.3	33.3
	6.7B	43.3	34.8	58.2	26.8	22.4	40.0	37.6
GPT-3†	13B	49.6	38.2	66.2	32.1	28.2	48.8	43.9
	175B	55.7	42.4	73.5	40.2	33.2	61.0	51.0
GPT-2 + PMI (Holtzman et al., 2021)	774M	43.9	38.3	47.0	31.6	43.2	44.5	41.4
	6.7B	48.5	39.8	51.5	33.0	48.0	50.3	45.2
GPT-3 + PMI† (Holtzman et al., 2021)	13B	**51.3**	42.1	57.7	38.5	50.4	58.5	49.8
	175B	55.7	43.7	63.3	45.5	58.0	66.7	55.5
Baselines (NLI format)								
Entailment (Yin et al., 2019)	355M	39.1	29.9	45.9	31.1	42.8	35.1	37.3
Ours (Sentence Encoder)								
RaLP$_{\text{SimCSE}}$	355M	39.4	35.1	48.3	26.5	38.2	47.3	39.1
RaLP$_{\text{ST5}}$	335M	40.5	38.3	56.2	28.7	44.4	55.5	43.9
RaLP$_{\text{ST5-XL}}$	1.24B	43.7	40.3	63.8	37.0	48.4	59.0	48.7
RaLP$_{\text{ST5-XXL}}$	4.8B	45.9	**42.3**	**69.3**	**44.5**	**52.8**	**63.9**	**53.1**

Table 3: Results for multiple choice tasks under zero-shot setting. The **Boldface** indicates the best performances in each column except the models with 175B parameters. The Underline shows the best scores in each column regardless of model size. We reproduce GPT-2 and GPT-2 + PMI with the code provided by Holtzman et al. (2021). The symbol † indicates the performance reported by Holtzman et al. (2021).

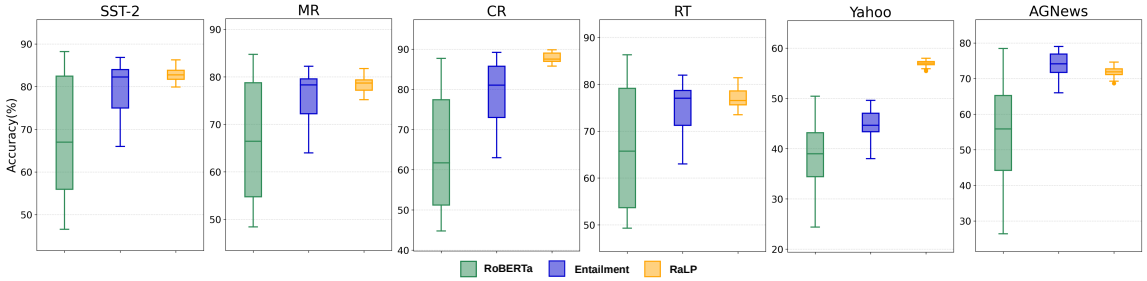


Figure 3: Results with varying verbalizers (6.1) in closed-set classification datasets. The middle line in each plot marks the mean score across all templates. Our approach has a smaller variance to template variations than Gao et al. (2021a); Yin et al. (2019).

ported (Jiang et al., 2021), cloze-style inference (RoBERTa) (Gao et al., 2021a) is overly sensitive to subtle contextual modifications. The Entailment (Yin et al., 2019) approach is less volatile, but it still suffers from sporadic performance drops with certain templates as shown in long lower whiskers in the plots. On the other hand, RaLP has more stable performance distribution under template variations, and this holds for all datasets.

6.2 Comparison with Few-shot Setup

If we allow *some* supervision (*i.e.* few-shot), there exist alternative strategies that can be used in combination with sentence encoder representations: linear probing and prototypical networks (Snell et al., 2017; Dopierre et al., 2021). Therefore, we figure out whether our zero-shot approach is still attractive as compared to the two few-shot options.

Baselines We constrain few-shot baselines to ensure minimal computational cost by keeping the backbone model unchanged. Linear probing requires training the classification head (*i.e.* logistic classification). The prototypical network treats the average embeddings of support examples as *class prototypes* and measures the distance of instance from these prototypes. For both methods, we randomly choose K=$\{2, 4, 8, 12, 16\}$ samples per class from training examples with 50 different seeds and report the averaged score.

Results Figure 4 illustrates that both few-shot learning baselines are volatile on sample selection variations. Specifically, linear probing exhibits more variability when given a small number of labeled samples. This may be because linear probing requires training parameters from scratch. Compared to few-shot training methods, RaLP shows robust performance when the supervision is ex-

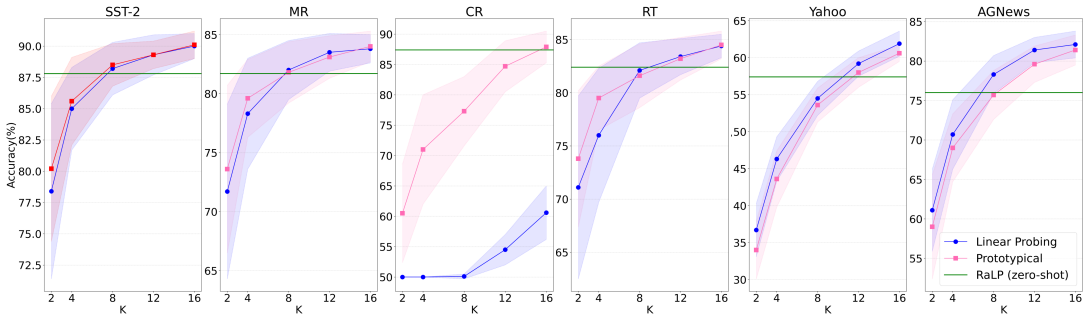


Figure 4: Performances of two few-shot approaches (mean and standard deviation across 50 runs) compared to RaLP. The x-axis represents the number of training examples per class. All use embeddings from *sentence-t5-large*.

tremely limited (i.e., $K \in [2, 4, 8]$).

7 Related Work

Prompting Brown et al. (2020) reformulate downstream tasks as language modeling with prompting, where the approach performs remarkably in few or zero-shot setup and even shows competitive performance to supervised methods (Schick and Schütze, 2021; Jiang et al., 2021). However, the performance of prompting methods varies greatly depending on which prompt is used (Perez et al., 2021; Lu et al., 2022). To address this issue, a line of works explore calibrating the predictions of language models for the few-shot inference (Holtzman et al., 2021; Zhao et al., 2021). Unlike previous prompting methods relying on language modeling (*i.e.* token prediction), our framework, which separately encodes label prompts and text input, is particularly robust to surface form variations.

Zero-shot Classification Prompting methods for zero-shot inference have been actively studied. These works use the likelihood of verbalizer given input text (Brown et al., 2020; Holtzman et al., 2021; Zhao et al., 2021), the conditional probability of the input text given verbalizer (Min et al., 2022a), and masked language model objective (Gao et al., 2021a) for solving NLU tasks. To facilitate zero-shot inference, Shi et al. (2022) incorporates information from an external corpus using a fuzzy verbalizer to adjust the decoding logits. NPM (Min et al., 2022c), on the other hand, removes the softmax function and only utilizes an external corpus to retrieve the answer.

Another line of research formulates the classification as a NLI task. Yin et al. (2019) solves the classification task by comparing the entailment logit scores where they concatenate the text input and label prompt as text pairs. Gera et al. (2022) further improves the framework by finetuning the model using self-training on unlabeled train data. SimPTC (Fei et al., 2022) proposes solving classification using sentence encoder and clustering method, which is similar to our approach without a retrieval component but differs in that it requires an in-domain unlabeled set.

Sentence Encoder Sentence encoders project the sentences to embedding space which can be applied to various language understanding tasks. (Reimers and Gurevych, 2019; Du et al., 2021; Gao et al., 2021b; Ni et al., 2022). Representations obtained by sentence encoders are well-suited features for classification in limited supervision setup (Tunstall et al., 2022). Existing methods based on sentence encoders train an additional classification head to solve downstream tasks, hindering them from using fine-grained representation and being unsuitable for zero-shot inference. In contrast, our approach does not incur additional parameters and fully exploits fine-grained representations.

8 Conclusion

This work empirically studies sentence encoders in a zero-shot text classification setup. The approach uses the label prompts as class anchors in an embedding space of a sentence encoder. To compensate for underspecified label names, we introduce RaLP, which retrieves the pseudo-label prompts from external knowledge sources and refines the prediction scores. We verify that RaLP is a powerful zero-shot classification strategy through an extensive evaluation of six closed-set text classification tasks and six open-set multiple-choice QA tasks. Further performance boost by retrieval demonstrates that class anchor selection is important and can be stably achieved by label prompt augmentation.

9 Limitations

Though RaLP demonstrates its strong performance in zero-shot classification tasks, the approach has several limitations. First, it is non-trivial to apply our method to non-classification tasks such as generation. Hence, future works can explore adapting the sentence encoders or their training methods in developing language models which have been explored in the context of the retrieval-augmented language models (Zhong et al., 2021). Additionally, RaLP relies on sentence encoders trained with the contrastive learning objective, which may introduce a certain level of dependency. Nonetheless, it is worth noting that publicly available NLI datasets can be used to obtain suitable sentence encoders as described in Appendix A. Moreover, while our work demonstrates its robust performance compared to some of the few-shot learning methods such as linear probing or prototypical network, we did not cover finetuning methods for our approach. Future works can explore tuning the parameters of the sentence encoders when given limited labeled data (Tunstall et al., 2022).

References

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. Raft: A real-world few-shot text classification benchmark. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. A neural few-shot text classification reality check. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yu Fei, Ping Nie, Zhao Meng, Roger Wattenhofer, and Mrinmaya Sachan. 2022. Beyond prompting: Making pre-trained language models better zero-shot learners by clustering representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics (TACL)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. Metaicl: Learning to learn in context. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wentau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022c. Nonparametric masked language modeling. *arXiv preprint arXiv:2212.01349*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multi-task learners. *Technical report*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. *arXiv preprint arXiv:2205.13792*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. volume 30.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2021. Training language models with memory augmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Appendix

A Analyzing Performance Gap between SimCSE and ST5

In Section 5.1, we evaluate the performance of various pre-trained sentence encoders using RaLP. Our results show that $\text{RaLP}_{\text{SimCSE}}$ (*sup-simcse-roberta-large*) has a lower performance than RaLP_{ST5} (*sentence-t5-large* although they have the same parameter size. We identify two main differences between SimCSE and ST5: i) the pre-trained model utilized as the initial checkpoint, and ii) the utilization of solely the NLI datasets in SimCSE as opposed to a combination of NLI and CommunityQA in ST5. Since the CommunityQA dataset is not publicly available, analyzing the reasons for this performance gap is important for future research on sentence encoders for zero-shot text classification. To do so, we conduct experiments by comparing the two encoder models.

For a fair comparison, we train the ST5 model using only the NLI datasets. Following the original setting in Ni et al. (2022), we employ T5-large (Raffel et al., 2020) as the initial pre-trained model and evaluate its performance through zero-shot text classification datasets. As shown in Table 4, the performance of the RaLP_{ST5} trained with the combination of CommunityQA and NLI achieved comparable performance with RaLP_{ST5} model trained using only the NLI dataset. However, among models trained using only NLI, RaLP_{ST5} outperforms $\text{RaLP}_{\text{SimCSE}}$ by 2.6 points on average. This suggests that the performance gap observed in Table 1 is primarily influenced by the T5 architecture rather than the CommunityQA dataset. We assume that the objective of reconstructing the consecutive span of corrupted tokens from one unique mask token (*i.e.* span masking) used in the pre-training of T5 has enhanced the model's general-purpose knowledge more than the objective of BERT-style (*i.e.* token masking).

B Evaluation Protocol

We report the accuracy for all datasets. To optimize hyperparameters, we exploit the validation dataset for each task. In case of CoQA, we sub-sample 3,000 instances from the training dataset, regarding them as in-house validation examples.

C Implementation Details

In all experiments, we evaluate models on a single A100 with 80GB, and A6000 GPU with 48GB of memory. We implement all models with PyTorch using sentence-transformers library from UKPLab[7]. We choose the best hyperparameter of top-K in $\{5, 10, 25, 50, 100\}$ for RaLP based on validation split.

[7] https://github.com/UKPLab/sentence-transformers

Method	Dataset	SST-2	MR	CR	RT	Yahoo	AGNews	Avg
RaLP_{ST5}	CommunityQA + NLI	**87.8**	**81.7**	87.4	**82.4**	57.4	76.6	**78.9**
RaLP_{ST5}	NLI	85.7	80.2	**87.8**	80.0	**58.8**	**76.8**	78.2
$\text{RaLP}_{\text{SimCSE}}$	NLI	82.3	78.0	87.1	76.7	57.0	72.6	75.6

Table 4: Results for ablation study of pre-training dataset. We additionally trained the ST5 model using only NLI dataset.

Dataset	**Verbalizers**
SST-2, MR, CR, RT	A `MASK` one.; It was `MASK`.; All in all `MASK`.; A `MASK` piece. (`MASK` = {great, terrible})
AGNews	Topic: `MASK`.; Subject: `MASK`.; This is about `MASK`.; It is about `MASK`. (`MASK` = {World, Sports, Business, Technology})
Yahoo	(Same as above) (`MASK` = {Company, Educational Institution, Artist, Athlete, Office Holder, Mean of Transportation, Building, Natural Place, Village, Animal, Plant, Album, Film, Written Work})

Table 5: The details of verbalizer setting in text classification task. We follow the verbalizers from Gao et al. (2021a); Min et al. (2022a)

Dataset	**Premise** x	**Hypothesis** y
RACE	There is not enough oil in the world now. As time goes by, it becomes less and less, so what are we going to do when it runs out[...].] question: According to the passage, which of the following statements is true?	answer: There is more petroleum than we can use now.
ARC	What carries oxygen throughout the body?	the answer is: red blood cells.
OBQA	Which of these would let the most heat travel through?	the answer is: a steel spoon in a cafeteria.
CoQA	Where can I stand on a river to see water falling without getting wet?	the answer is: bridge.

Table 6: The details of verbalizer setting in multiple choice datasets. The hypothesis candidates are directly given as label prompt in each task. The texts with blue color denote templates that we prepend following Holtzman et al. (2021)

.

Dataset	Verbalizers
SST-2, MR, CR, RT	It was MASK.; A MASK piece.; A MASK one.; All in all MASK.; It was absolutely MASK, really.; It was really so MASK.; It was more than just MASK.; And that was absolutely MASK, too.; It was still pretty MASK.; It was all MASK.; It was literally MASK.; It's a MASK thing.; What a MASK performance.; It is an MASK piece at the best of times.; One MASK piece.; A MASK work.; A MASK play.; This is a MASK one.; It is utterly MASK.; A really MASK one.; That's MASK.; That's all MASK.; All told this is a truly MASK thing.; A MASK overall.; All together MASK. (MASK = {great, terrible})
AGNews	Topic: MASK.; Subject: MASK.; This is about MASK.; It is about MASK.; It's about the MASK.; It's about MASK.; It's all about the MASK.; It's just about the MASK.; It's the whole lot with the MASK.; Here, we are talking about the MASK.; This is the theme of the MASK.; It is related to the MASK.; It is about what it means for the MASK.; This involves the MASK.; Theme: MASK.; keyword: MASK.; On a related topic: the MASK.; It is for the MASK.; The subject: the MASK.; Main topic: MASK.; Content: MASK.; Theme is the MASK.; Issue: MASK.; Executive Summary: MASK.; Material: MASK. (MASK = World, Sports, Business, Technology)
Yahoo	(Same as above) (MASK = {Company, Educational Institution, Artist, Athlete, Office Holder, Mean of Transportation, Building, Natural Place, Village, Animal, Plant, Album, Film, Written Work})

Table 7: We use all listed verbalizers in verbalizer variation test in Section 6.1. We evaluate on total 25 verbalizer with or without punctuation.

Original Label Prompt	Retrieval Augmented Label Prompts
It was great.	great, love it - great, love it - great, love it - great, love it really did love it and it was also great ;
	It was really good! I reallly liked it! As does my father and mother and brother and sister-in-law. It was epic. ;
	i loved it. it could not have been better. justin kirby ;
	it was awesome, maiden is at there best, is was totally awesomeness, nobody comes close to maiden, it's a totally fulfillment ;
	it was awesome!!!!! of course it was awesome. its Harry Potter for lord's sake!!!!
It was good.	A <unk> . " A is good " , or " A was good " . ;
	great, love it - great, love it - great, love it - great, love it really did love it and it was also great ;
	It was ok. Wish I could give this a great review, sorry. But, it was still ok, if you're interested. ;
	It was alright having not seen it for many years i enjoyed watching it and am glad for the purchase ;
	It was really good! I reallly liked it! As does my father and mother and brother and sister-in-law. It was epic.
It was famous.	= = = In popular memory = = = ;
	= = Famous people who witnessed it = = ;
	= = = National and international fame = = = ;
	This is a real classic, with a number of actors and actresses that were to become very famous later on. ;
	This classic was very innovative for it's time and is a classic.The Princess created quite a stir at the time, and a cult following.
It was terrible.	terrible, terrible.It wasm't worth the money or the time to watch it. we turned it off in the middle of it ;
	It was horrible. One of the worst books ever. Thank youBelly Up ;
	This book was terrible and was horrific when it came to the description. ;
	It was awful. I went into it with great hopes expecting to enjoy it. I wanted to like it, I wanted to laugh. I spent more time rolling my eyes and checking the counter to see how much time was left till it was over. I was so disappointed. ;
	Just poor quality and I just didn't like it.Really couldn't pay attention because it was terrible, awful movie I wouldn't recommend.
It was horrible.	It was horrible. One of the worst books ever. Thank youBelly Up ;
	This book was terrible and was horrific when it came to the description. ;
	It was sick, depressing, annoying, and not to mention rotten! so that's what I have to say on the subject! ;
	I hated it and thought it was really bad. It was twisted and is really scary to litte children ;
	truth be told, this is a horrible story of what really happened - so very awful...it was impossible to read
It was awful.	It was horrible. One of the worst books ever. Thank youBelly Up ;
	It was awful. I went into it with great hopes expecting to enjoy it. I wanted to like it, I wanted to laugh. I spent more time rolling my eyes and checking the counter to see how much time was left till it was over. I was so disappointed. ;
	This book was terrible and was horrific when it came to the description. ;
	It was sick, depressing, annoying, and not to mention rotten! so that's what I have to say on the subject! ;
	The movie was truly a wretched experience. The six letters to describe it are S-U-C-K-E-D!

Table 8: Original label prompt in sentiment analysis task (SST-2, MR, CR, and RT) and retrieval augmented label prompt in Section 5.1.

Original Label Prompt	Retrieval Augmented Label Prompts
Topic: world.	= = = = Worldwide = = = = ; = = = = Informing the world = = = = ; = = = On the Wisdom of this World = = = ; = State of the World (song) = ; = = = = HELLO , WORLD = = = =
Topic: international.	= = = = International = = = = ; = = = = International affairs = = = = ; = = = = International services = = = = ; = = = = International relations = = = = ; = = = International (InLine) = = =
Topic: global.	= = = = Worldwide = = = = ; = = = Global environmental impact = = = ; = = = From the global community = = = ; = = = Around the globe = = = ; = = = = Terrestrial globes = = = =
Topic: sports.	= = = = Sports = = = = ; = = = = Athletic history = = = = ; = = = Effects on sports = = = ; = = = World sport context = = = ; = = = = Sport = = = =
Topic: entertainment.	= = = = Entertainment = = = = ; = = = Entertainment and culture = = = ; Entertainment is a form of activity that holds the attention and interest of an audience , or gives pleasure and delight . It can be [...] and even for a global audience . ; = = = Sports and entertainment = = = ; = = = Media and entertainment = = =
Topic: recreation.	= = = = Recreations = = = = ; = = = = Recreation = = = = ; = = = Sports and recreation = = = ; = = = Sport and recreation = = = ; = = = = Tourism and recreation = = = =
Topic: business.	= = = = Business = = = = ; = = = = On labor and business = = = = ; = = = Lines of business = = = ; = = = DE 1 Business = = = ; = = = Operations and business = = =
Topic: economics.	= = = = Economic lecture = = = = ; = = = = Economics = = = = ; = = = Views on economics = = = ; = = = = Economic = = = = ; = = = Philosophy of economics = = =
Topic: financial.	= = = = Finances = = = = ; = = = Financial and loan = = = ; = = = = Fiscal issues = = = = ; = = = = Financial performance = = = = ; = = = Economy and finance = = =
Topic: technology.	= = = Influence on technology = = = ; = = = Views on the technology = = = ; = = = = Technical aspects = = = = ; = = = Effects of technology = = = ; = = = Technology and science = = =
Topic: science.	Science – news on science @-@ related topics (e.g. cool technology , space telescope observations , interesting medical research) . ; = = = Interactions with the scientific community = = = ; = = = = Scientific uses = = = = ; = = = History of science = = = ; = = = Science and scientism = = =
Topic: mathmatics.	= = = Attention towards mathematics = = = ; = = = = Mathematics = = = = ; = = = An Introduction to Mathematics = = = ; = = = Mathematics as an art = = = ; = = = Philosophy of mathematics = = =

Table 9: Original label prompt for AGNews and retrieval augmented label prompt in Section 5.1.