Perfect: Prompt-free and Efficient Language Model Fine-Tuning

Anonymous ACL submission

Abstract

Current methods for few-shot fine-tuning of pretrained masked language model (PLM) require carefully engineered prompts and verbalizers for each new task, to convert examples into a cloze-format that the PLM can score. In this work, we propose Perfect, a simple and efficient method for few-shot fine-tuning of PLMs without relying on any such handcrafting, which is highly effective given as few as 32 data points. Perfect makes two key design choices: First, we show that manually engineered task prompts can be replaced with task-specific adapters that enable sample-efficient fine-tuning and reduce memory and storage costs by roughly factors of 5 and 100, respectively. Second, instead of using handcrafted verbalizers, we learn a new multi-token label embedding during fine-tuning which are not tied to the model vocabulary and which allow us to avoid complex auto-regressive decoding. These embeddings are not only learnable from limited data but also enable nearly 100x faster training and inference. Experiments on a wide range of few shot NLP tasks demonstrate that Perfect, while being simple and efficient, also outperforms existing state-of-the-art few-shot learning methods.1

1 We will release our code publicly to facilitate future work.

1 Introduction

Recent methods for few-shot language model fine-tuning obtain impressive performance but require careful engineering of prompts and verbalizers to convert inputs to a cloze-format (Taylor, 1953) that can be scored with pre-trained language models (PLMs) (Radford et al., 2018; Radford et al.; Brown et al., 2020; Schick and Schütze, 2021a,b). For example, as Figure 1 shows, a sentiment classifier can be designed by inserting the input text $x$ in a prompt template “$x$ It was [MASK]” where verbalizers (e.g., ‘great’ and ‘terrible’) are substituted for the [MASK] to score target task labels (‘positive’ or ‘negative’). In this paper, we show that such engineering is not needed for few-shot learning and instead can be replaced with simple methods for data-efficient fine-tuning with as few as 32 end-task examples.

More specifically, we propose Perfect, a Prompt-free and Efficient paradigm for FEw-shot Cloze-based fine-Tuning. To remove handcrafted verbalizers, Perfect uses task-specific adapter layers (Houlsby et al., 2019) (§3.1). Freezing the underlying PLM with millions or billions of parameters (Liu et al., 2019; Raffel et al., 2020), and only tuning adapters with very few new parameters saves on memory and storage costs (§4.2), while allowing very sample-efficient tuning (§4). It also stabilizes the training by increasing the worst-case performance and decreasing variance across the choice of examples in the few shot training sets (§4.3).

To remove handcrafted verbalizers (with variable token lengths), we introduce a new multi-token fixed-length classifier scheme that learns task label embeddings which are independent from the language model vocabulary during fine-tuning (§3.2). We show (§4) that this approach is sample efficient and outperforms carefully engineered verbalizers from random initialization (§4). It also allows us to avoid previously used expensive auto-regressive decoding schemes (Schick and Schütze, 2021b), by leveraging prototypical networks (Snell et al., 2017) over multiple tokens. Overall, these changes enable up to 100x faster learning and inference (§4.2).

Overall, Perfect has several advantages: It avoids engineering patterns and verbalizers for each...
new task, which can be cumbersome. Recent work has shown that even some intentionally irrelevant or misleading prompts can perform as well as more interpretable ones (Webson and Pavlick, 2021). Unlike the zero-shot or extreme few-shot case, where prompting might be essential, we argue in this paper that all you need is tens of training examples to avoid these challenges by adopting PERFECT or a similar data-efficient learning method. Experiments on a wide variety of NLP tasks demonstrate that PERFECT outperforms state-of-the-art prompt-based methods while being significantly more efficient in inference and training time, storage, and memory usage (§4.2).

To the best of our knowledge, we are the first to propose a few-shot learning method with PLMs that successfully removes all per-task manual engineering.

2 Background

Problem formulation: We consider a general problem of fine-tuning language models in a few-shot setting, on a small training set with \( K \) unique classes and \( N \) examples per class, such that the total number of examples is \( |D| = N \times K \). Let \( D = \bigcup_{k=1}^{K} D_k \) be the given training set, where \( D_k = \{(x_i^k, y_i^k)\}_{i=1}^N \) shows the set of examples labeled with class \( k \) and \( y_i^k \in \mathcal{Y} \) is the corresponding label, where \( |\mathcal{Y}| = K \). We additionally assume access to a development set with the same size as the training data. Note that larger validation sets can grant a substantial advantage (Perez et al., 2021), and thus it is important to use a limited validation size to be in line with the goal of few-shot learning. Unless specified otherwise, in this work, we use 16 training examples (\( N = 16 \)) and a validation set with 16 examples, for a total of 32-shot learning.

2.1 Adapters

Recent work has shown that fine-tuning all parameters of PLMs with a large number of parameters in low-resource datasets can lead to a suboptimal solution (Peters et al., 2019; Dodge et al., 2020). As shown in Figure 2, Rebuffi et al. (2018) and Houlsby et al. (2019) suggest an efficient alternative, by inserting small task-specific modules called adapters within layers of a PLMs. They then only train the newly added adapters and layer normalization, while fixing the remaining parameters of a PLM.

Each layer of a transformer model is composed of two primary modules: a) an attention block, and b) a feed-forward block, where both modules are followed by a skip connection. As depicted in Figure 2, adapters are normally inserted after each of these blocks before the skip connection.

Figure 2: Left: Adapter integration in a PLM. Right: An adapter architecture. Adapters are usually inserted after the feed-forward and self-attention modules. During training, we only optimize the green components.

Adapters are bottleneck architectures. By keeping input and output dimensions the same, they introduce no additional architectural changes. Each adapter, \( A(.) \in \mathbb{R}^{H} \), consists of a down-projection, \( D(.) \in \mathbb{R}^{H \times B} \), a non-linearity, such as GeLU (Hendrycks and Gimpel, 2016), and an up-projection \( U(.) \in \mathbb{R}^{B \times H} \), where \( H \) is the dimension of input hidden states \( x \), and \( B \) is the bottleneck size. Formally defined as:

\[
A(x) = U(\text{ReLU}(D(x))) + x, \tag{1}
\]

2.2 Prompt-based Fine-tuning

Standard Fine-tuning: In standard fine-tuning with PLMs (Devlin et al., 2019), first a special [CLS] token is appended to the input \( x \), and then the PLM maps it to a sequence of hidden representations \( h = (h_1, \ldots, h_S) \) with \( h_i \in \mathbb{R}^H \), where \( H \) is the hidden dimension, and \( S \) is the maximum sequence length. Then, a classifier, \( \text{softmax}(W^T h_{\text{CLS}}) \), using the embedding of the classification token (\( h_{\text{CLS}} \)), is trained end-to-end for each downstream task. The main drawback of this approach is the discrepancy between the pre-training and fine-tuning phases since PLMs have been trained to predict mask tokens in a masked language modeling task (Devlin et al., 2019).

Prompt-based tuning: To address this discrepancy, prompt-based fine-tuning (Schick and Schütze, 2021a,b; Gao et al., 2021) formulates tasks in a cloze-format (Taylor, 1953). This way, the model can predict targets with a masked language modeling (MLM) objective. For example, as shown in Figure 1, for a sentiment classification task, inputs are converted to:

\[ x_{\text{prompt}} = [\text{CLS}] \cdot x \cdot \underbrace{[\text{MASK}] \cdot [\text{SEP}]}_{\text{pattern}} \]
Then, the PLM determines which verbalizer (e.g., ‘great’ and ‘terrible’) is the most likely substitute for the mask in the $x_{\text{prompt}}$. This subsequently determines the score of targets (‘positive’ or ‘negative’). In detail:

Training strategy: Let $M : \mathcal{Y} \rightarrow \mathcal{V}$ be a mapping from target labels to individual words in a PLM’s vocabulary. We refer to this mapping as verbalizers. Then the input is converted to $x_{\text{prompt}} = T(x)$ by appending a pattern and a mask token to $x$ so that it has the format of a masked language modeling input. Then, the classification task is converted to a MLM objective (Tam et al., 2021; Schick and Schütze, 2021a), and the PLM computes the probability of the label $y$ as:

$$p(y|x) = p([\text{MASK}] = M(y)|x_{\text{prompt}}) = \frac{\exp(W^T_M y_M h_{\text{[MASK]}})}{\sum_{v' \in \mathcal{V}} \exp(W^T_v h_{\text{[MASK]}})}, \quad (2)$$

where $h_{\text{[MASK]}}$ is the last hidden representation of the mask, and $W_v$ shows the output embedding of the PLM for each verbalizer $v \in \mathcal{V}$. For many tasks, verbalizers have multiple tokens. Schick and Schütze (2021b) extended (2) to multiple mask tokens by adding the maximum number of mask tokens $M$ needed to express the outputs (verbalizers) for a task. In that case, Schick and Schütze (2021b) computes the probability of each class as the summation of the log probabilities of each token in the corresponding verbalizer, and then they add a hinge loss to ensure a margin between the correct verbalizer and the incorrect ones.

Inference strategy: During inference, the model needs to select which verbalizer to use in the given context. Schick and Schütze (2021b) predicts the verbalizer tokens in an autoregressive fashion. They first trim the number of mask tokens from $M$ to each candidate verbalizer’s token length, and compute the probability of each mask token. Then they choose the predicted token with the highest probability and replace the corresponding mask token. Conditioning on this new token, the probabilities of the remaining mask positions are recomputed. They repeat this autoregressive decoding until they fill all mask positions. This inference strategy is very slow, as the number of forward passes increases with the number of classes and the number of verbalizer’s tokens.

This formulation obtained impressive few-shot performance with PLMs. However, the success of this approach heavily relies on engineering handcrafted patterns and verbalizers. Coming up with suitable verbalizers and patterns can be difficult (Mishra et al., 2021). Additionally, the performance is sensitive to the wording of patterns (Zhao et al., 2021; Perez et al., 2021; Schick and Schütze, 2021a; Jiang et al., 2020) or to the chosen verbalizers (Webson and Pavlick, 2021).

In addition, handcrafted verbalizers cause problems for efficient training: a) they require updating the PLM embedding layer, causing large memory overhead; b) fine-tuning PLMs also requires a very small learning rate (usually $10^{-5}$), which slows down tuning the parameters of the verbalizers; c) modeling verbalizers as one of the tokens of the PLM vocabulary (perhaps unintentionally) impacts the input representation during tuning; d) verbalizers have variable token lengths, complicating the implementation in a vectorized format, thereby making it challenging to efficiently fine-tune PLMs.

3 Method

We propose PERFECT, a verbalizer and patterns-free few-shot learning method. We design PERFECT to be close to the pre-training phase, similar to the PET family of models (Schick and Schütze, 2021b; Gao et al., 2021), while replacing handcrafted patterns and
verbalizers with new components that are designed to describe the task and learn the labels. As shown in Figure 3, we first convert each input $x_{\text{input}}$ to its masked language modeling (MLM) input containing $M$ mask tokens $[\text{MASK}]$ with no added patterns, denoted as $x_{\text{masked}} = \mathcal{T}(x_{\text{input}})$.\textsuperscript{3} PERFECT then trains a classifier per-token and optimizes the average multi-class hinge loss over each mask position.

Three main components play a role in the success of PERFECT: a) a pattern-free task description, where we use task-specific adapters to efficiently tell the model about the given task, replacing previously manually engineered patterns (§3.1), b) multi-token label-embedding as an efficient mechanism to learn the label representations, removing manually designed verbalizers (§3.2), c) an efficient inference strategy building on top of the idea of prototypical networks (Snell et al., 2017) (§3.4), which replaces prior iterative autoregressive decoding methods (Schick and Schütze, 2021b).

As shown in Figure 3, we fix the underlying PLM model and only optimize the new parameters that we add (green boxes). This includes the task-specific adapters to adapt the representations for a given task and the multi-token label representations. We detail each of these components below.

### 3.1 Pattern-Free Task Description

We use task-specific adapter layers to provide the model with learned, implicit task descriptions. Adapters additionally bring multiple other benefits: a) fine-tuning all weights of PLMs with millions or billions of parameters is sample-inefficient, and can be unstable in low-resource settings (Dodge et al., 2020); adapters allow sample-efficient fine-tuning, by keeping the underlying PLM fixed, b) adapters reduce the storage and memory footprints (§4.2), c) they also increase stability and performance (§4), making them an excellent choice for few-shot fine-tuning. To our knowledge, this is the first approach for using task-specific adapters to effectively and efficiently remove patterns in few-shot learning. Experimental results in §4 show its effectiveness compared to handcrafted patterns and soft prompts (Li and Liang, 2021; Lester et al., 2021).

\textsuperscript{3}We discuss the general case with inserting multiple masks; for some datasets this improves performance (§4.3.1).

\textsuperscript{4}We insert mask tokens after the input string in single-sentence benchmarks, and after the first sentence in the case of sentence-pair datasets and encode both sentences as a single input, which we found to perform the best (Appendix C).

### 3.2 Multi-Token Label Embeddings

We freeze the weights of the PLM’s embedding layer and introduce a separate label embedding $L \in \mathbb{R}^{K \times M \times H}$, which is a multi-token label representation where $M$ is the number of tokens representing each label, $K$ indicates the number of classes, $H$ is the input hidden dimension. Using a fixed number of tokens $M$ for each label, versus variable-token length verbalizers used in prior work (Schick and Schütze, 2021a,b) substantially simplifies the implementation and accelerates the training (§4.2).

### 3.3 Training PERFECT

As shown in Figure 3, we optimize label embeddings so that the PLM predicts the correct label, and optimize adapters to adapt the PLM for the given task. For label embeddings, PERFECT trains a classifier per token and optimizes the average multi-class hinge loss over all mask positions. Given $x_{\text{masked}}$, let $h_{[\text{MASK}]}$ be the embedding of its $i$-th mask token from the last layer of the PLM encoder. Additionally, let $f(\cdot) : \mathbb{R}^H \rightarrow \mathbb{R}^K$ be a per-token classifier that computes the predictions by multiplying the mask token embedding with its corresponding label embedding. Formally defined as:

$$t_i = f(h_{[\text{MASK}]}),$$

where $L_i \in \mathbb{R}^{K \times H}$ shows the label embedding for the $i$-th mask position. Then, for each mask position, we optimize a multi-class hinge loss between their scores $t_i$ and labels. Formally defined as:

$$L(x, y, i) = \sum_{k=1, k \neq y}^{K} \max(0, m - t_{ik} + t_{iy}),$$

where $t_{ik}$ shows the $k$-th element of $t_i$, representing the score corresponding to class $k$, and $m$ is the margin, which we fix to the default value of $m = 1$. Then, the final loss is computed by averaging the loss over all mask tokens and training samples:

$$L = \frac{1}{M|D|} \sum_{(x, y) \in D} \sum_{i=1}^{M} L(x, y, i).$$

### 3.4 Inference with PERFECT

During evaluation, instead of relying on the prior iterative autoregressive decoding schemes (Schick and Schütze, 2021b), we classify a query point by looking at the score corresponding to class $k$, and taking the class with the highest score. For label embeddings, we classify a query point by looking at the score corresponding to class $k$, and taking the class with the highest score.
where $d$ is squared euclidean distance, $h_i^Q$ indicates the embedding of the $i$-th mask position for the query sample $q$, and $c_{iy} \in \mathbb{R}^D$ is the prototype representation of the $i$-th mask token with class label $y$, i.e., the mean embedding of $i$-th mask position in all training samples with label $y$: 

$$c_{iy} = \frac{1}{|D_y|} \sum_{b \in D_y} h_i^b,$$

where $h_i^b$ shows the embedding of $i$-th mask position for training sample $b$, and $D_y$ is the training instances with class $y$. This strategy closely follows prototypical networks (Snell et al., 2017), but applied across multiple tokens. We choose this form of inference because prototypical networks are known to be sample efficient and robust (Snell et al., 2017), and because it substantially speeds up evaluation compared to prior methods (§4.2).

4 Experiments

We conduct extensive experiments on a variety of NLP datasets to evaluate the performance of PERFECT and compare it with state-of-the-art few-shot learning.

Datasets: We consider 7 tasks and 12 datasets: 1) the sentiment analysis datasets SST-2 (Socher et al., 2013), SST-5 (Socher et al., 2013), MR (Pang and Lee, 2005), and CR (Hu and Liu, 2004), 2) the subjectivity classification dataset SUBJ (Pang and Lee, 2004), 3) the question classification dataset TREC (Voorhees and Tice, 2000), 4) the natural language inference datasets CB (De Marneffe et al., 2019) and RTE (Wang et al., 2019a), 5) the question answering dataset QNLI (Rajpurkar et al., 2016), 6) the word sense disambiguation dataset WiC (Pilehvar and Camacho-Collados, 2019), 7) the paraphrase detection datasets MRPC (Dolan and Brockett, 2005) and QQP. See datasets statistics in Appendix A.

For MR, CR, SST-5, SUBJ, and TREC, we test on the original test sets, while for other datasets, since test sets are not publicly available, we test on the original validation set. We sample 16 instances per label from the training set to form training and validation sets.

Baselines We compare with the state-of-the-art few-shot learning of PET and fine-tuning:

PET (Schick and Schütze, 2021a,b) is the state-of-the-art few-shot learning method that employs carefully crafted verbalizers and patterns. We report the best (PET-best) and average (PET-average) results among all patterns and verbalizers.6

FINETUNE The standard fine-tuning (Devlin et al., 2019), with adding a classifier on top of the [CLS] token and fine-tuning all parameters.

Our method We study the performance of PERFECT and perform an extensive ablation study to show the effectiveness of our design choices:

PERFECT-rand We randomly initialize the label embedding $L$ from a normal distribution $\mathcal{N}(0,\sigma)$ with $\sigma = 10^{-4}$ (chosen based on validation performance, see Appendix D) without relying on any handcrafted patterns and verbalizers. As an ablation, we study the following two variants:

PERFECT-init We initialize the label embedding with the token embeddings of manually designed verbalizers in the PLM’s vocabulary to study the impact of engineered verbalizers.

PERFECT-prompt We compare using adapters versus soft prompt-tuning with removing adapters and appending trainable continuous prompt embeddings to the input (Lester et al., 2021). Then, we only tune the soft prompt and the label embedding.

Experimental details: We use the RoBERTa large model (Liu et al., 2019) (355M parameters) as the underlying PLM for all methods. We use the HuggingFace PyTorch implementation (Wolf et al., 2020). For the baselines, we used the carefully manually designed patterns and verbalizers in Gao et al. (2021), Min et al. (2021), and Schick and Schütze (2021b) (usually 5 different options per dataset; see Appendix B).

We evaluate all methods using 5 different random samples to create the training/validation sets and 4 different random seeds for training. Therefore, for PET-average, we report the results on 20 x 5 (number of patterns and verbalizers) = 100 runs, while for PET-best and our method, we report the results over 20 runs. The variance in few-shot learning methods is usually high (Perez et al., 2021; Zhao et al., 2021; Lu et al., 2021). Therefore, we report average, worst-case performance, and standard deviation across all runs, where the last two values can be important for risk-sensitive applications (Asri et al., 2016).

4.1 Experimental Results

Table 1 shows the performance of all methods. PERFECT obtains state-of-the-art results, improving the performance compared to PET-average by +1.1

For a controlled study, we use the MLM variant shown in (2), which has been shown to perform the best (Tam et al., 2021).

---

6We also tried with cosine similarity but found a slight improvement with squared Euclidean distance (Snell et al., 2017).

5https://quoradata.quora.com/
Table 1: Performance of all methods on single-sentence and sentence-pair benchmarks. We report average/worst-case accuracy/standard deviation. PERFECT obtains the state-of-the-art results. Bold fonts indicate the best results.

<table>
<thead>
<tr>
<th>Method</th>
<th>SST-2</th>
<th>CR</th>
<th>MR</th>
<th>SST-5</th>
<th>Subj</th>
<th>TREC</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-Sentence Benchmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINETUNE</td>
<td>81.4/70.9/4.0</td>
<td>80.1/72.3/4.1</td>
<td>77.7/76.8/4.6</td>
<td>39.2/34.3/2.5</td>
<td>90.2/84.1/1.8</td>
<td>87.6/75.8/3.7</td>
<td>76.0/67.3/4.4</td>
</tr>
<tr>
<td>PET-Average</td>
<td>89.7/81.0/2.4</td>
<td>88.4/86.8/3.0</td>
<td>85.9/79.0/2.1</td>
<td>45.9/40.3/2.4</td>
<td>88.1/79.6/2.4</td>
<td>85.0/70.6/4.5</td>
<td>80.5/69.9/2.8</td>
</tr>
<tr>
<td>PET-Best</td>
<td>89.1/81.9/2.6</td>
<td>88.8/85.8/1.9</td>
<td>86.4/82.0/1.5</td>
<td>46.0/41.2/2.4</td>
<td>88.7/84.6/1.8</td>
<td>85.8/70.6/4.4</td>
<td>80.8/74.2/2.4</td>
</tr>
<tr>
<td>PERFECT-rand</td>
<td>90.7/86.2/1.2</td>
<td>86.3/81.4/1.6</td>
<td>42.7/35.1/2.9</td>
<td>89.1/82.8/2.1</td>
<td>90.6/81.6/3.2</td>
<td>91.6/85.8/2.1</td>
<td></td>
</tr>
<tr>
<td>PERFECT-init</td>
<td>90.9/87.6/1.5</td>
<td>89.7/87.4/1.2</td>
<td>85.4/75.8/3.3</td>
<td>42.8/35.9/3.5</td>
<td>87.6/81.6/2.8</td>
<td>90.4/86.1/4.1</td>
<td>81.1/75.8/2.4</td>
</tr>
<tr>
<td>PERFECT-prompt</td>
<td>70.6/56.9/3.1</td>
<td>71.0/55.8/2.2</td>
<td>66.6/49.6/3.3</td>
<td>32.2/26.5/3.2</td>
<td>82.7/76.6/3.9</td>
<td>79.6/66.8/5.0</td>
<td>67.1/54.6/6.2</td>
</tr>
<tr>
<td>Ablation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>CB</td>
<td>RTE</td>
<td>QNLI</td>
<td>MRPC</td>
<td>QQP</td>
<td>WiC</td>
<td>Avg</td>
</tr>
<tr>
<td>Sentence-Pair Benchmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINETUNE</td>
<td>72.9/67.9/2.5</td>
<td>56.8/50.2/3.8</td>
<td>62.7/51.4/7.0</td>
<td>70.1/62.7/4.7</td>
<td>65.0/59.8/3.6</td>
<td>52.4/46.1/3.7</td>
<td>63.3/56.4/4.2</td>
</tr>
<tr>
<td>PET-Average</td>
<td>86.9/73.2/5.1</td>
<td>60.1/49.5/4.7</td>
<td>66.5/55.7/6.2</td>
<td>62.1/38.2/6.8</td>
<td>63.4/44.7/7.9</td>
<td>51.0/36.1/6.2</td>
<td>65.0/51.2/5.6</td>
</tr>
<tr>
<td>PET-Best</td>
<td>90.0/78.6/9.7</td>
<td>62.3/51.8/4.5</td>
<td>70.5/57.9/8.4</td>
<td>63.4/49.3/6.5</td>
<td>70.7/55.2/5.8</td>
<td>51.6/47.2/2.5</td>
<td>68.1/56.6/4.9</td>
</tr>
<tr>
<td>PERFECT-rand</td>
<td>90.3/83.9/3.5</td>
<td>60.4/53.1/4.7</td>
<td>74.1/60.3/4.6</td>
<td>67.8/54.7/5.7</td>
<td>71.2/64.2/3.5</td>
<td>53.8/47.0/3.0</td>
<td>69.6/60.5/4.2</td>
</tr>
<tr>
<td>Ablation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>CB</td>
<td>RTE</td>
<td>QNLI</td>
<td>MRPC</td>
<td>QQP</td>
<td>WiC</td>
<td>Avg</td>
</tr>
</tbody>
</table>

and +4.6 points for single-sentence and sentence-pair datasets respectively. It even outperforms PET-best, where we report the best performance of PET across multiple manually engineered patterns and verbalizers. Moreover, PERFECT improves the minimum performance and reduces standard deviation substantially. Finally, PERFECT is also significantly more efficient: reducing the training and inference time, memory usage, and storage costs (see §4.2).

PET-best improves the results over PET-average showing that PET is unstable to the choice of patterns and verbalizers; this difference is more highlighted for sentence-pair benchmarks. This can be because the position of the mask highly impacts the results, and patterns used in sentence-pair datasets in Schick and Schütze (2021b) well leverages this by putting the mask in multiple locations (see Appendix B).

As an ablation, even if we initialize the label embedding with handcrafted verbalizers, PERFECT-init consistently obtains lower performance, demonstrating that PERFECT is able to obtain state-of-the-art performance with learning from pure random initialization. We argue that initializing randomly close to zero (with low variance $\sigma = 10^{-5}$), as done in our case, slightly improves performance, which perhaps is not satisfied when initializing from the manually engineered verbalizers (see Appendix D).

As a second ablation, when learning patterns with optimizing soft prompts in PERFECT-prompt, we observe high sensitivity to learning rate, as also confirmed in Li and Liang (2021) and Mahabadi et al. (2021a). We experimented with multiple learning rates but performance consistently lags behind PERFECT-rand. This can be explained by the low flexibility of such methods as all the information regarding specifying patterns needs to be contained in the prefixes. As a result, the method only allows limited interaction with the rest of the model parameters, and obtaining good performance requires very large models (Lester et al., 2021). In addition, increasing the sequence length leads to memory overhead (Mahabadi et al., 2021a), and the number of prompt tokens is capped by the number of tokens that can fit in the maximum input length, which can be a limitation for tasks requiring large contexts.

### 4.2 Efficiency Evaluation

In this section, we compare the efficiency of PERFECT with the state-of-the-art few-shot learning method, PET. To this end, we train all methods for ten epochs on the 500-sampled QNLI dataset. We select the largest batch size for each method that fits a fixed
Due to the auto-regressive inference strategy of PET (Schick and Schütze, 2021b), every prior work implemented it with a batch size of 1 (Perez et al., 2021; Schick and Schütze, 2021b; Tam et al., 2021). Additionally, since PET deals with verbalizers of variable lengths, it is hard to implement their training phase in a batch mode. We specifically choose QNLI to have verbalizers of the same length and enable batching for comparison purposes (referred to as PET in batch). However, verbalizers are still not of the fixed-length for most other tasks, and this speed-up does not apply generally to PET.

In Table 2, we report the percentage of trained parameters, memory usage of each method, training, and inference time. PERFECT reduces the number of trained parameters, and therefore the storage requirement, by 99.08%. It additionally reduces the memory requirement by 21.93% compared to PET. PERFECT speeds up training substantially, by 97.22% relative to the original PET’s implementation, and 30.85% to our implementation of PET. This is because adapter-based tuning saves on memory and allows training with larger batch sizes. On the other hand, PERFECT is significantly faster during inference time (96.76% less inference time relative to PET).

Overall, given the size of PLMs with millions and billions of parameters (Liu et al., 2019; Raffel et al., 2020), efficient few-shot learning methods are of paramount importance for practical applications. PERFECT not only outperforms the state-of-the-art in terms of accuracy and stability (Table 1), but also is significantly more efficient in runtime, storage, and memory.

### 4.3 Analysis

**Can task-specific adapters replace manually engineered patterns?** PERFECT is a pattern-free approach and employs adapters to provide the PLMs with task descriptions implicitly. In this section, we study the contribution of replacing manual patterns with adapters in isolation without considering our other contributions in representing labels, training, and inference. In PET (Schick and Schütze, 2021a,b), we replace the handcrafted patterns with task-specific adapters (Pattern-Free) while keeping the verbalizers and the training and inference intact\(^7\) and train it with a similar setup as in §4. Table 3 shows the results. While PET is very sensitive to the choice of prompts, adapters provide an efficient alternative to learn patterns robustly by improving the performance (average and worst-case) and reducing the standard deviation. This finding demonstrates that task-specific adapters can effectively replace manually engineered prompts. Additionally, they also save on the training budget by at least 1/number of patterns (normally 1/5) by not requiring running the method for different choices of patterns, and by freezing most parameters, this saves on memory and offers additional speed-up.

#### 4.3.1 Ablation Study

**Impact of Removing Adapters** To study the impact of adapters in learning patterns, we remove adapters, while keeping the label embedding. Handcrafted patterns are not included and we tune all parameters of the model. Table 4 shows the results. Adding adapters for learning patterns contributes to the performance by improving the average performance, and making the model robust by improving the minimum performance and reducing the standard deviation. This is because training PLMs with millions of parameters is sample-inefficient\(^7\)

---

\(^7\)Since we don’t have patterns, in the case of multiple sets of verbalizers, we use the first set of verbalizers as a random choice.

---

### Table 2: Percentage of trained parameters, average peak memory, training, and inference time. ∆% is the relative difference with respect to PET. Lower is better.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PET</th>
<th>PERFECT</th>
<th>∆%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained params (M)</td>
<td>355.41</td>
<td>3.28</td>
<td>-99.08%</td>
</tr>
<tr>
<td>Peak memory (GB)</td>
<td>20.93</td>
<td>16.34</td>
<td>-21.93%</td>
</tr>
<tr>
<td>Training time (min)</td>
<td>23.42</td>
<td>0.65</td>
<td>-97.22%</td>
</tr>
<tr>
<td>+ PET in batch</td>
<td>0.94</td>
<td>0.65</td>
<td>-30.85%</td>
</tr>
<tr>
<td>Inference time (min)</td>
<td>9.57</td>
<td>0.31</td>
<td>-96.76%</td>
</tr>
</tbody>
</table>

### Table 3: Average performance of PET with five different datasets vs. Pattern-Free that replaces handcrafted patterns with task-specific adapters. We report the average/worst-case performance and the standard deviation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PET-Average</th>
<th>Pattern-Free</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>89.7/81.0/2.4</td>
<td>90.5/87.8/1.2</td>
</tr>
<tr>
<td>CR</td>
<td>88.4/68.8/3.0</td>
<td>89.8/87.0/1.4</td>
</tr>
<tr>
<td>MR</td>
<td>85.9/79.0/2.1</td>
<td>86.4/83.0/1.8</td>
</tr>
<tr>
<td>SST-5</td>
<td>45.9/40.3/2.4</td>
<td>44.8/40.0/2.4</td>
</tr>
<tr>
<td>SUBJ</td>
<td>88.1/79.6/2.4</td>
<td>85.3/74.7/3.8</td>
</tr>
<tr>
<td>TREC</td>
<td>85.0/70.6/4.5</td>
<td>87.9/84.6/1.8</td>
</tr>
<tr>
<td>CB</td>
<td>86.9/73.2/5.1</td>
<td>93.0/89.3/1.9</td>
</tr>
<tr>
<td>RTE</td>
<td>60.1/49.5/4.7</td>
<td>63.7/56.3/4.1</td>
</tr>
<tr>
<td>QNLI</td>
<td>66.5/55.7/6.2</td>
<td>71.3/65.8/2.5</td>
</tr>
<tr>
<td>MRPC</td>
<td>62.1/38.2/6.8</td>
<td>66.0/54.4/5.6</td>
</tr>
<tr>
<td>QQP</td>
<td>63.4/44.7/7.9</td>
<td>71.8/64.3/3.7</td>
</tr>
<tr>
<td>WiC</td>
<td>51.0/46.1/2.6</td>
<td>53.7/50.3/2.0</td>
</tr>
</tbody>
</table>

### Table 4: Average performance of PET with five different datasets vs. Pattern-Free that replaces handcrafted patterns with task-specific adapters. We report the average/worst-case performance and the standard deviation.
We report the average performance/worst-case performance/and the standard deviation.

Table 4: Performance of \texttt{PERFECT} w/o adapters, \texttt{-Adapters}.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>\textbf{PERFECT}</th>
<th>\textbf{-Adapters}</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>90.7/88.2/1.2</td>
<td>88.2/81.9/2.3</td>
</tr>
<tr>
<td>CR</td>
<td>90.0/85.5/1.4</td>
<td>89.2/83.1/1.7</td>
</tr>
<tr>
<td>MR</td>
<td>86.3/81.4/1.6</td>
<td>82.5/78.2/2.5</td>
</tr>
<tr>
<td>SST-5</td>
<td>42.7/35.1/2.9</td>
<td>40.6/33.6/3.3</td>
</tr>
<tr>
<td>SUBJ</td>
<td>89.1/82.8/2.1</td>
<td>89.7/85.0/1.9</td>
</tr>
<tr>
<td>TREC</td>
<td>90.6/81.6/3.2</td>
<td>89.8/74.2/4.3</td>
</tr>
<tr>
<td>CB</td>
<td>90.3/83.9/3.5</td>
<td>89.6/83.9/2.8</td>
</tr>
<tr>
<td>RTE</td>
<td>60.4/53.1/4.7</td>
<td>61.7/53.8/5.1</td>
</tr>
<tr>
<td>QNLI</td>
<td>74.1/60.3/4.6</td>
<td>73.2/56.3/5.8</td>
</tr>
<tr>
<td>MRPC</td>
<td>67.8/54.7/5.7</td>
<td>68.0/54.2/6.1</td>
</tr>
<tr>
<td>QQP</td>
<td>71.2/64.2/3.5</td>
<td>71.0/62.0/3.7</td>
</tr>
<tr>
<td>WiC</td>
<td>53.8/47.0/3.0</td>
<td>52.5/46.9/3.0</td>
</tr>
<tr>
<td>Avg</td>
<td>75.6/68.1/3.1</td>
<td>74.7/66.1/3.5</td>
</tr>
</tbody>
</table>

Table 4: Performance of \texttt{PERFECT} w/o adapters, \texttt{-Adapters}. We report the average performance/worst-case performance and the standard deviation.

and unstable on resource-limited datasets (Dodge et al., 2020; Zhang et al., 2020). However, by using adapters, we substantially reduce the number of trainable parameters, allowing the model to be better tuned in a few-shot setting.

**Impact of the number of masks** In Table 1, to compare our design with PET in isolation, we fixed the number of mask tokens as the maximum number inserted by PET. In Table 5, we study the impact of varying the number of inserted mask tokens for a random selection of six tasks. For most tasks, having two mask tokens performs the best, while for MR and RTE, having one, and for MRPC, inserting ten masks improves the results substantially. The number of required masks might be correlated with the difficulty of tasks. \texttt{PERFECT} is designed to be general, enabling having multiple mask tokens.

5 Related Work

**Adapter Layers:** Mahabadi et al. (2021b) and Üstün et al. (2020) proposed to generate adapters’ weights using hypernetworks (Ha et al., 2017), where Mahabadi et al. (2021b) proposed to share a small hypernetwork to generate conditional adapter weights efficiently for each transformer layer and task. Mahabadi et al. (2021a) proposed compacter layers by building on top of ideas of parameterized hypercomplex layers (Zhang et al., 2021) and low-rank methods (Li et al., 2018; Aghajanyan et al., 2021), as an efficient fine-tuning method for PLMs. We are the first to employ adapters to replace handcrafted patterns for few-shot learning.

**Few-shot Learning with PLMs:** Researchers continuously tried to address the challenges of manually engineered patterns and verbalizers: a) Learning the patterns in a continuous space (Li and Liang, 2021; Qin and Eisner, 2021; Lester et al., 2021), while freezing PLM for efficiency, has the problem that, in most cases, such an approach only works with very large scale PLMs (Lester et al., 2021), and lags behind full fine-tuning in a general setting, while being inefficient and not as effective compared to adapters (Mahabadi et al., 2021a). b) Optimizing patterns in a discrete space (Shin et al., 2020; Jiang et al., 2020; Gao et al., 2021) has the problem that such methods are computationally costly. c) Automatically finding verbalizers in a discrete way (Schick et al., 2020; Schick and Schütze, 2021a) is computationally expensive and does not perform as well as manually designed ones. d) Removing manually designed patterns (Logan IV et al., 2021) substantially lags behind the expert-designed ones. Our proposed method, \texttt{PERFECT}, does not rely on any handcrafted patterns and verbalizers.

6 Conclusion

We proposed \texttt{PERFECT}, a simple and efficient method for few-shot learning with pre-trained language models without relying on handcrafted patterns and verbalizers. \texttt{PERFECT} employs task-specific adapters to learn task descriptions implicitly, replacing previous handcrafted patterns, and a continuous multi-token label embedding to represent the output classes. Through extensive experiments over 12 NLP benchmarks, we demonstrate that \texttt{PERFECT}, despite being far simpler and more efficient than recent few-shot learning methods, produces state-of-the-art results. Overall, the simplicity and effectiveness of \texttt{PERFECT} make it a promising approach for few-shot learning with PLMs.
References


Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. ACL.


David Ha, Andrew Dai, and Quoc V. Le. 2017. Hypernetworks. In ICLR.


Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In SIGKDD.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? TACL.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. EMNLP.


Chunyuan Li, Heeraad Forkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In ICLR.


Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021b. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *ACL.*


Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL.*

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *NeurIPS.*

Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *RepL4NLP.*


Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP.*


Tim Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *COLING.*

Tim Schick and Hinrich Schütze. 2021a. Exploiting close-questions for few-shot text classification and natural language inference. In *EACL.*

Tim Schick and Hinrich Schütze. 2021b. It’s not just size that matters: Small language models are also few-shot learners. In *NAACL.*


Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP.*


Aston Zhang, Yi Tay, SHUAI Zhang, Alvin Chan, Anh Tuan Luu, Siu Hui, and Jie Fu. 2021. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with 1/n parameters. In ICLR.


Table 6: Statistics of datasets used in this work. We sample $N \times |\mathcal{Y}|$ instances (with multiple seeds) from the original training set to form the few-shot training and validation sets. The test column shows the size of the test set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>#Train</th>
<th>#Test</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>Sentiment analysis</td>
<td>6920</td>
<td>872</td>
<td>2</td>
</tr>
<tr>
<td>SST-5</td>
<td>Sentiment analysis</td>
<td>8544</td>
<td>2210</td>
<td>5</td>
</tr>
<tr>
<td>SUBJ</td>
<td>Subjectivity classification</td>
<td>8000</td>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>TREC</td>
<td>Question classification</td>
<td>5452</td>
<td>500</td>
<td>6</td>
</tr>
</tbody>
</table>

**Single-Sentence Benchmarks**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>#Train</th>
<th>#Test</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>Sentiment analysis</td>
<td>8662</td>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>CR</td>
<td>Sentiment analysis</td>
<td>1774</td>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>MRPC</td>
<td>Paraphrase detection</td>
<td>3668</td>
<td>408</td>
<td>2</td>
</tr>
<tr>
<td>QNLI</td>
<td>Question answering</td>
<td>104743</td>
<td>5463</td>
<td>2</td>
</tr>
<tr>
<td>QQP</td>
<td>Paraphrase detection</td>
<td>363846</td>
<td>40430</td>
<td>2</td>
</tr>
</tbody>
</table>

**Sentence-Pair Benchmarks**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>#Train</th>
<th>#Test</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>Natural language inference</td>
<td>250</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>RTE</td>
<td>Natural language inference</td>
<td>2490</td>
<td>277</td>
<td>2</td>
</tr>
<tr>
<td>WiC</td>
<td>Word sense disambiguation</td>
<td>5428</td>
<td>638</td>
<td>2</td>
</tr>
<tr>
<td>MRPC</td>
<td>Paraphrase detection</td>
<td>3668</td>
<td>408</td>
<td>2</td>
</tr>
<tr>
<td>QNLI</td>
<td>Question answering</td>
<td>104743</td>
<td>5463</td>
<td>2</td>
</tr>
<tr>
<td>QQP</td>
<td>Paraphrase detection</td>
<td>363846</td>
<td>40430</td>
<td>2</td>
</tr>
</tbody>
</table>

A Experimental Details

**Datasets** Table 6 shows the statistics of the datasets used. We download SST-2, MR, CR, SST-5, and SUBJ from Gao et al. (2021), while the rest of the datasets are downloaded from the HuggingFace Datasets library (Lhoest et al., 2021b,a). RTE, CB, WiC datasets are from SuperGLUE benchmark (Wang et al., 2019a), while QQP, MRPC and QNLI are from GLUE benchmark (Wang et al., 2019b) with Creative Commons license (CC BY 4.0). RTE (Wang et al., 2019a) is a combination of data from RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giammichele et al., 2007), and RTE5 (Bentivogli et al., 2009). For WiC (Pilehvar and Camacho-Collados, 2019) sentences are selected from VerbNet (Schuler, 2005), WordNet (Miller, 1995), and Wiktionary.

**Computing infrastructure** We run all the experiments on one NVIDIA A100 with 40G of memory.

**Training hyper-parameters** We set the maximum sequence length based on the recommended values in the HuggingFace repository (Wolf et al., 2020) and prior work (Min et al., 2021; Schick and Schütze, 2021b), i.e., we set it to 256 for SUBJ, CR, CB, RTE, and WiC, and 128 for other datasets. For all methods, we use a batch size of 32. For FINETUNE and PET, we use the default learning rate of $10^{-5}$, while for our method, as required by adapter-based methods (Mahabadi et al., 2021a), we set the learning rate to a higher value of $10^{-4}$. Through all experiments, we fix the adapter bottleneck size to 64. Following Pfeiffer et al. (2021), we experimented with keeping one of the adapters in each layer for better training efficiency and found keeping the adapter after the feed-forward module in each layer to perform the best. For tuning label embedding, we use the learning rate of \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\} and choose the one obtaining the highest validation performance. For PERFECT-prompt, we tune the continuous prompt for learning rate of \{10^{-1}, 10^{-2}, 10^{-3}\}. Following Lester et al. (2021), for PERFECT-prompt, we set the number of prompt tokens to 20, and initialize them with a random subset of the top 5000 token’s embedding of the PLM. We train all methods for 6000 steps. Based on our results, this is sufficient to allow the models to converge. We save a checkpoint every 100 steps for all methods and report the results for the hyper-parameters performing the best on the validation set for each task.

B Choice of Patterns and Verbalizers

For SST-2, MR, CR, SST-5, and TREC, we used 4 different patterns and verbalizers from Gao et al. (2021). For CB, WiC, RTE datasets, we used the designed patterns and verbalizers in Schick and Schütze (2021b). For QQP, MRPC, and QNLI, we wrote the patterns and verbalizers inspired by the ones in Schick and Schütze (2021b). The used patterns and verbalizers are as follows:

- For sentiment analysis tasks (MR, CR, SST-2, SST-5), given a sentence $s$:

  $s$ A <MASK> one.

  $s$ It was <MASK>.

  $s$ All in all <MASK>.

  $s$ A <MASK> piece.

  with "great" as a verbalizer for positive, "terrible" for negative. In case of SST-5 with five labels, we expand it to "great", "good", "okay", "bad", and "terrible".

  \(^8\)We have also tried to tune the baselines with the learning rate of $10^{-4}$ but it performed worst.

  \(^9\)We also tried tuning prompts with learning rates of \{10^{-4}, 10^{-5}\} but it performed worst, as also observed in prior work (Mahabadi et al., 2021a; Min et al., 2021).
• For **SUBJ**, given a sentence $s$:

- $s$ This is <MASK>.
- $s$ It's all <MASK>.
- $s$ It's <MASK>.
- $s$ Is it <MASK>?

with "subjective" and "objective" as verbalizers.

• For **TREC**, given a question $q$, the task is to classify the type of it:

- $q$ <MASK>:
- $q$ Q:<MASK>:
- $q$ why <MASK>?
- $q$ Answer: <MASK>.

with "Description", "Entity", "Expression", "Human", "Location", "Number" as verbalizers for question types of "Description", "Entity", "Abbreviation", "Human", "Location", and "Numeric".

• For entailment task (**RTE**) given a premise $p$ and hypothesis $h$:

- "$h$ ? | <MASK>, "p"
- $h$? | <MASK>, p
- "$h$ ? | <MASK>, p

with "Yes" as a verbalizer for entailment, "No" for contradiction, "Maybe" for neutral.

• For entailment task (**QNLI**), given a sentence $s$ and question $q$:

- "$h$ ? | <MASK>, "p"
- $h$? | <MASK>, p
- "$h$ ? | <MASK>, p

Based on the previous sentence, $q$? <MASK>.

with "Yes" or "true" as verbalizers for entailment and "No" or "false" for not entailment.

• For entailment task (**CB**), given a premise $p$ and hypothesis $h$:

- "$h$ ? | <MASK>, "p"
- $h$? | <MASK>, p
- "$h$ ? | <MASK>, p

with "Yes" as a verbalizer for entailment, "No" for contradiction.

- $p$ question: $h$ True or False? answer: <MASK>.

with "true" as a verbalizer for entailment, "false" for contradiction, "neither" for neutral.

• For **QQP**, given two questions $q_1$ and $q_2$:

- Do $q_1$ and $q_2$ have the same meaning?<MASK>.
- $q_1$. Based on the previous question, $q_2$? <MASK>.

with "Yes" or "true" as verbalizers for duplicate and "No" or "false" for not duplicate.

- Based on the following question, $q_1$?<MASK>.

with "Yes" and "No" as verbalizers for entailment and not entailment respectively.

• For **QNLI** and **CB**, given a premise $p$ and hypothesis $h$:

- "$h$ ? | <MASK>, "p"
- $h$? | <MASK>, p
- "$h$ ? | <MASK>, p

with "Yes" as a verbalizer for entailment, "No" for contradiction.

- $p$ question: $h$ True or False? answer: <MASK>.

with "true" as a verbalizer for entailment, "false" for contradiction.

- Based on the following question, $q_1$? <MASK>.

with "Yes" and "No" as verbalizers for duplicate and not duplicate respectively.
• For MRPC, given two sentences $s_1$ and $s_2$:

Do $s_1$ and $s_2$ have the same meaning? <MASK>.

with "Yes" or "true" as verbalizers for equivalent and "No" or "false" for not equivalent.

$s_1$. Based on the previous sentence, $s_2$? <MASK>.

Based on the following sentence, $s_1$? <MASK>$s_2$.

with "Yes" and "No" as verbalizers for equivalent and not equivalent respectively.

• For WiC, given two sentences $s_1$ and $s_2$ and a word $w$, the task is to classify whether $w$ is used in the same sense.

"$s_1$" / "$s_2$". Similar sense of "$w$"? <MASK>.

$s_1$, $s_2$. Does $w$ have the same meaning in both sentences? <MASK>.

With "No" and "Yes" as verbalizers for False, and True.

"w". Sense (1) (a) "$s_1$" (<MASK>) "$s_2$".

With "2" and "b" as verbalizers for False, and True.

### C Impact of the Position of Masks in Sentence-pair Datasets

We evaluate the impact of the position of mask tokens in sentence-pair benchmarks. Given two sentences $s_1$ and $s_2$, we consider the following four locations for inserting mask tokens, where in the case of encoding as two sentences, input parts to the encoder are separated with $|$:

1. $s_1$ $s_2$ <MASK>
2. $s_1$ <MASK>$s_2$
3. $s_1$ | <MASK>$s_2$
4. $s_1$ $s_2$<MASK>

### D Impact of Initialization

We initialize the label embedding matrix with random initialization from a normal distribution $\mathcal{N}(0, \sigma)$. In table 8, we show the development results for different values of $\sigma$. We choose the $\sigma$ obtaining the highest performance on average over average and worst case performance, i.e., $\sigma = 10^{-4}$.