

# It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners

Timo Schick<sup>1,2</sup> and Hinrich Schütze<sup>1</sup>

<sup>1</sup> Center for Information and Language Processing, LMU Munich, Germany

<sup>2</sup> Sulzer GmbH, Munich, Germany

timo.schick@sulzer.de

## Abstract

When scaled to hundreds of billions of parameters, pretrained language models such as GPT-3 (Brown et al., 2020) achieve remarkable few-shot performance on challenging natural language understanding benchmarks. In this work, we show that performance similar to GPT-3 can be obtained with language models whose parameter count is several orders of magnitude smaller. This is achieved by converting textual inputs into cloze questions that contain some form of task description, combined with gradient-based optimization; additionally exploiting unlabeled data gives further improvements. Based on our findings, we identify several key factors required for successful natural language understanding with small language models.<sup>1</sup>

## 1 Introduction

Pretraining ever-larger language models (LMs) on massive plain text corpora has led to significant improvements on a wide range of NLP tasks (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2019, *inter alia*). A standard approach is to replace the pretrained model’s output layer with a task-specific head and finetune the entire model on a set of labeled training data. However, language modeling is not only a powerful pretraining objective, but many tasks can be reformulated as cloze questions (e.g., by appending phrases such as “the correct answer is \_\_\_”), allowing pretrained LMs to solve them without any or with only very few labeled examples (Radford et al., 2019; Schick and Schütze, 2020a).

Very recently, Brown et al. (2020) introduced GPT-3, a pretrained LM with an enormous 175 billion parameters, and showed that it has amazing few-shot abilities: By reformulating tasks

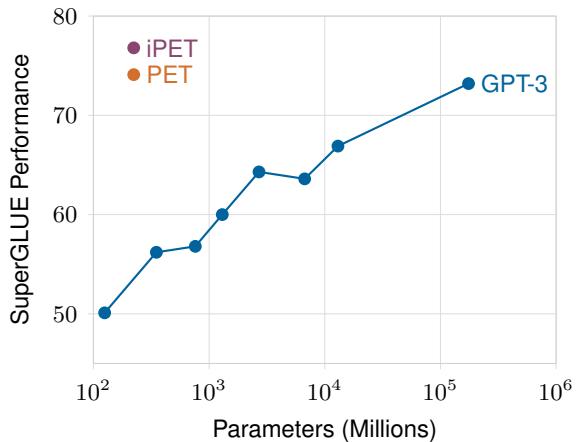


Figure 1: Performance of ALBERT with PET/iPET and GPT-3 on SuperGLUE for 32 training examples. ALBERT with PET/iPET outperforms GPT-3 even though it has three orders of magnitude fewer parameters.

as language modeling problems, GPT-3 achieves near state-of-the-art results for some tasks in the SuperGLUE benchmark (Wang et al., 2019) given just 32 labeled examples. This is achieved through priming: GPT-3 is given a few demonstrations of inputs and corresponding outputs as context for its predictions, but no gradient updates are performed. While being straightforward to use, this method has two major drawbacks:

- It requires a gigantic LM to work well, making it unusable in many real-world scenarios.
- It does not scale to more than a few examples as the context window of most LMs is limited to a few hundred tokens.

An alternative to priming is *pattern-exploiting training* (PET) (Schick and Schütze, 2020a), which combines the idea of reformulating tasks as cloze questions with regular gradient-based finetuning. While PET additionally requires unlabeled data, unlabeled data is much easier to obtain than labeled

<sup>1</sup>Our implementation is publicly available at <https://github.com/timoschick/pet>.

examples for many real-world applications. Crucially, PET only works when the answers to be predicted by the LM correspond to a single token in its vocabulary; this is a severe limitation as many tasks cannot easily be worded that way.

In this work, we modify PET to also work for tasks that require predicting more than one token. We then show that in combination with ALBERT (Lan et al., 2020), PET and its iterative variant (iPET) both outperform GPT-3 on SuperGLUE with 32 training examples, while requiring only 0.1% of its parameters (Figure 1). Finally, we show that similar performance can also be achieved without unlabeled data and provide a detailed analysis of the factors contributing to PET’s strong performance: its ability to combine multiple task formulations, its resilience to wordings that are hard to understand, its usage of labeled data, and characteristics of the underlying LM.

## 2 Related Work

Enabling LMs to perform zero-shot learning by providing task descriptions was proposed by Radford et al. (2019) and has been applied to text classification (Puri and Catanzaro, 2019), commonsense knowledge mining (Davison et al., 2019) and argumentative relation classification (Opitz, 2019). It is also commonly used for probing the knowledge contained within LMs (Trinh and Le, 2018; Petroni et al., 2019; Talmor et al., 2019; Schick and Schütze, 2020b; Ettinger, 2020, *i.a.*).

As finding ways to reformulate tasks as cloze questions that are understood well by LMs is difficult (Jiang et al., 2019), Schick and Schütze (2020a) propose PET, a method that uses knowledge distillation (Hinton et al., 2015) to easily combine several reformulations. In contrast to PET, which uses gradient-based optimization, Radford et al. (2019) and Brown et al. (2020) investigate priming, where examples are given as context but no parameter updates are performed.

Our modified version of PET uses masked language models (Devlin et al., 2019) to assign probabilities to sequences of text; this is similar to using them in a generative fashion (Wang and Cho, 2019) and has previously been investigated by Salazar et al. (2020) and Ghazvininejad et al. (2019).

## 3 Pattern-Exploiting Training

Let  $M$  be a masked language model (MLM),  $T$  its vocabulary and  $\_\_ \in T$  the mask token. For some

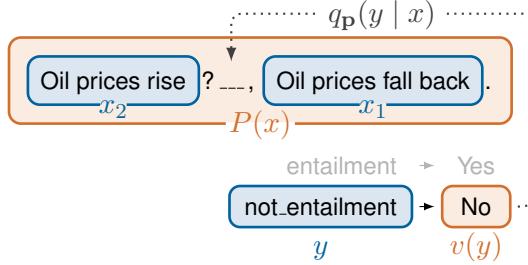


Figure 2: Application of a PVP  $\mathbf{p} = (P, v)$  for recognizing textual entailment: An input  $x = (x_1, x_2)$  is converted into a cloze question  $P(x)$ ;  $q_{\mathbf{p}}(y | x)$  for each  $y$  is derived from the probability of  $v(y)$  being a plausible choice for the masked position.

$\mathbf{z} \in T^*$  containing at least  $k$  masks and  $t \in T$ , we denote with  $q_M^k(t | \mathbf{z})$  the probability that  $M$  assigns to  $t$  at the  $k$ th masked position in  $\mathbf{z}$ ; the model’s logits before applying softmax are denoted with  $s_M^k(t | \mathbf{z})$ . We consider the task of mapping inputs  $x \in X$  to outputs  $y \in Y$ , for which PET requires a set of *pattern-verbalizer pairs* (PVPs). Each PVP  $\mathbf{p}$  consists of

- a *pattern*  $P : X \rightarrow T^*$  that maps inputs to cloze questions containing a single mask;
- a *verbalizer*  $v : Y \rightarrow T$  that maps each output to a single token representing its task-specific meaning in the pattern.

As illustrated in Figure 2, the core idea of PET is to derive the probability of  $y$  being the correct output for  $x$  from the probability of  $v(y)$  being the “correct” token at the masked position in  $P(x)$ . Based on this intuition, a conditional probability distribution  $q_{\mathbf{p}}$  of  $Y$  given  $X$  is defined as

$$q_{\mathbf{p}}(y | x) = \frac{\exp s_{\mathbf{p}}(y | x)}{\sum_{y' \in Y} \exp s_{\mathbf{p}}(y' | x)} \quad (1)$$

where  $s_{\mathbf{p}}(y | x) = s_M^1(v(y) | P(x))$  is the raw score of  $v(y)$  at the masked position in  $P(x)$ .

For a given task, identifying PVPs that perform well is challenging in the absence of a large development set. Therefore, PET enables a combination of multiple PVPs  $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$  as follows:

1. For each PVP  $\mathbf{p}$ , a MLM is finetuned on training examples  $(x, y)$  by minimizing the cross entropy between  $y$  and  $q_{\mathbf{p}}(y | x)$ . In practice, Schick and Schütze (2020a) train three MLMs per pattern as performance can vary substantially between runs.

2. The ensemble of finetuned MLMs is used to annotate a set of unlabeled examples; each unlabeled example  $x \in X$  is annotated with soft labels based on the probability distribution

$$q_{\mathbf{P}}(y | x) \propto \exp \sum_{\mathbf{p} \in \mathbf{P}} w_{\mathbf{p}} \cdot s_{\mathbf{p}}(y | x) \quad (2)$$

similar to Eq. 1 where  $w_{\mathbf{p}}$  is a weighting term that is proportional to the accuracy achieved with  $\mathbf{p}$  on the training set *before* training.

3. The resulting soft-labeled dataset is used to train a regular sequence classifier by minimizing cross entropy between its output and  $q_{\mathbf{P}}$ .

As steps (2) and (3) above closely resemble knowledge distillation (Hinton et al., 2015), we also refer to them simply as *distillation*.

To give MLMs trained on different patterns further opportunity to learn from one another, Schick and Schütze (2020a) also propose iPET, an iterative variant of PET in which several generations of models are trained on datasets of increasing size that are labeled by previous generations. This is achieved as follows: First, an ensemble of MLMs is trained as in regular PET. For each model  $M_i$ , a random subset of other models is used to generate a new training set  $T_i$  by assigning labels to those unlabeled examples for which the selected subset of models is most confident in its prediction. Each  $M_i$  is then retrained on  $T_i$ ; this process is repeated several times, each time increasing the number of examples in  $T_i$  by a constant factor.

### 3.1 PET with Multiple Masks

An important limitation of PET is that the verbalizer  $v$  must map each possible output to a *single* token, which is impractical or outright impossible for many tasks. We thus generalize verbalizers to functions  $v : Y \rightarrow T^*$ ; this requires some modifications to inference and training.<sup>2</sup> We further generalize PET in that we do not assume the output space to be identical for each input: for each  $x \in X$ , we denote with  $Y_x \subseteq Y$  the set of possible outputs given  $x$  as input. Given a PVP  $\mathbf{p} = (P, v)$ , we define  $l(x) = \max_{y \in Y_x} |v(y)|$  to be the maximum number of tokens required to express any output in  $Y_x$  and  $P^k(x)$  to be  $P(x)$  with the mask token replaced by  $k$  masks.

<sup>2</sup>While our approach can easily be adapted to models capable of doing conditional generation with bidirectional context (e.g., Lewis et al., 2020; Raffel et al., 2019), we stick with MLMs as they are more lightweight and we found them to perform better on simple cloze tasks in preliminary experiments.

**Inference** For  $x \in X$ ,  $y \in Y_x$  and  $|v(y)| = k$ , we redefine  $q_{\mathbf{p}}(y | x)$  in an autoregressive fashion: Starting from  $P^k(x)$ , we perform  $k$  consecutive predictions, where we always select the next token to predict based on the MLM’s confidence. That is, we set  $q_{\mathbf{p}}(y | x) = q(v(y) | P^k(x))$  where

$$q(t_1 \dots t_k | \mathbf{z}) = \begin{cases} 1 & \text{if } k = 0 \\ q_M^j(t_j | \mathbf{z}) \cdot q(t' | \mathbf{z}') & \text{if } k \geq 1 \end{cases} \quad (3)$$

with  $j = \arg \max_{i=1}^k q_M^j(t_j | \mathbf{z})$ ,  $\mathbf{z}'$  is  $\mathbf{z}$  except  $\mathbf{z}'_j = t_j$  and  $t' = t_1 \dots t_{j-1} t_{j+1} \dots t_k$ . Note that unlike in original PET (Eq. 1),  $q_{\mathbf{p}}$  is not a probability distribution as its values do not sum to one.

**Training** Computing  $q_{\mathbf{p}}(y | x)$  as in Eq. 3 for each training example  $(x, y)$  would be prohibitively expensive. To enable computation of all required probabilities in a single forward pass, we approximate  $q_{\mathbf{p}}(y | x)$  by (i) always inserting the maximum number of mask tokens required to express any output and (ii) for each  $y' \in Y_x$ , predicting all tokens in  $v(y') = t_1 \dots t_k$  in parallel, where we simply ignore the model’s predictions for all  $l(x) - k$  superfluous mask tokens:

$$\tilde{q}_{\mathbf{p}}(y' | x) = \prod_{i=1}^k \tilde{q}_M^i(t_i | P^{l(x)}(x)) \quad (4)$$

As  $\tilde{q}_{\mathbf{p}}$  is not a probability distribution over  $Y_x$ , cross entropy is not an ideal training objective as it can also be minimized by reducing the probability assigned to sequences  $\mathbf{z} \notin v(Y_x)$  that are not part of the output space, despite this having no effect on the model’s prediction. We instead opt for multi-class hinge loss (Weston and Watkins, 1999; Dogan et al., 2016) and minimize:

$$\sum_{y' \in Y_x} \max(0; 1 - \log \tilde{q}_{\mathbf{p}}(y | x) + \log \tilde{q}_{\mathbf{p}}(y' | x)) \quad (5)$$

That is, we require the difference between the log probability of  $y$  and the log probability of any output  $y' \in Y_x \setminus \{y\}$  to be at least 1.

## 4 Experiments

We use the SuperGLUE benchmark (Wang et al., 2019) to compare PET and GPT-3. We cannot evaluate PET using the exact same training data as GPT-3 because for most tasks, GPT-3 uses a different set of training examples for each test example and for the other tasks, training sets were not available upon request; however, the exact choice

of examples has little impact on GPT-3’s performance.<sup>3</sup> We thus create new training sets by randomly selecting 32 examples for each task using a fixed random seed. Following previous work (Raffel et al., 2019; Brown et al., 2020), we select only positive examples for WSC; for both MultiRC and ReCoRD, we follow Brown et al. (2020) and select a total of 32 questions.

We additionally create sets of up to 20,000 unlabeled examples for each task; this is done by removing all labels from the original training sets. As the training sets for RTE and CB are very small, we additionally select random unlabeled examples from MNLI (Williams et al., 2018) for both tasks. We refer to the resulting sets of training examples and unlabeled examples as *FewGLUE*.<sup>4</sup>

#### 4.1 Tasks

Below, we describe each of the SuperGLUE tasks and our corresponding PVPs. We use a vertical bar (|) to mark boundaries between text segments.

**BoolQ** (Clark et al., 2019) is a QA task where each example consists of a passage  $p$  and a yes/no question  $q$ . We use the following patterns:

- $p$ . Question:  $q$ ? Answer: \_\_\_
- $p$ . Based on the previous passage,  $q$ ? \_\_\_
- Based on the following passage,  $q$ ? \_\_\_  $p$

We define two verbalizers mapping questions containing a true statement to yes/true and others to no/false, respectively, for a total of 6 PVPs.

**CB** (De Marneffe et al., 2019) and **RTE** (Dagan et al., 2006) are textual entailment tasks like MNLI, so we use PVPs similar to Schick and Schütze (2020a). For a premise  $p$  and hypothesis  $h$ , we use

$h?|...p$  , “ $h?$ ” $|...p$  ,  $h?|...p$  , “ $h?$ ” $|...p$

and a verbalizer that maps entailment to yes, disagreement to no and neutral to maybe.

Given a premise  $p$ , the task in **COPA** (Roemheld et al., 2011) is to determine the *cause* or *effect* of the premise given two options  $c_1$  and  $c_2$ . For determining the *effect*, we use the following patterns:

“ $c_1$ ” or “ $c_2$ ”?  $p$ , so \_\_\_ ,  $c_1$  or  $c_2$ ?  $p$ , so \_\_\_

<sup>3</sup>Based on personal correspondence with the authors.

<sup>4</sup>The FewGLUE dataset is publicly available at <https://github.com/timoschick/fewglue>.

For determining the *cause*, we use the same patterns but replace so with because. The verbalizer for  $c_1$  and  $c_2$  is the identity function.

For **WiC** (Pilehvar and Camacho-Collados, 2019), given a word  $w$  and two sentences  $s_1$  and  $s_2$  in which it occurs, the task is to decide if  $w$  is used with the same sense in both sentences. We use:

- “ $s_1$ ” / “ $s_2$ ”. Similar sense of “ $w$ ”? \_\_\_
- $s_1 s_2$  Does  $w$  have the same meaning in both sentences? \_\_\_
- $w$ . Sense (1) (a) “ $s_1$ ” (...) “ $s_2$ ”

For the first two patterns, we use yes as verbalization for words used in the same sense and no for other words; for the third pattern, we use b and 2.

For **WSC** (Levesque et al., 2011), each example consists of a sentence  $s$  with a marked pronoun  $p$  and noun  $n$ , and the task is to determine whether  $p$  refers to  $n$ . We follow previous work (Raffel et al., 2019; Brown et al., 2020) and treat WSC as a generative task. We highlight  $p$  in  $s$  by putting it in asterisks and use the following patterns:

- $s$  The pronoun ‘\* $p$ \*’ refers to \_\_\_
- $s$  In the previous sentence, the pronoun ‘\* $p$ \*’ refers to \_\_\_
- $s$  In the passage above, what does the pronoun ‘\* $p$ \*’ refer to? Answer: \_\_\_

We use the identity function as verbalizer for  $n$ . Note that WSC is different from other tasks in that it requires free-form completion. This in turn requires some modifications during training and inference that are discussed in Appendix A.

**MultiRC** (Khashabi et al., 2018) is a QA task. Given a passage  $p$ , a question  $q$  and an answer candidate  $a$ , the task is to decide whether  $a$  is a correct answer for  $q$ . We use the same verbalizer as for BoolQ and similar patterns:

- $p$ . Question:  $q$ ? Is it  $a$ ? \_\_\_
- $p$ . Question:  $q$ ? Is the correct answer “ $a$ ”? \_\_\_
- $p$ . Based on the previous passage,  $q$ ? Is “ $a$ ” a correct answer? \_\_\_

For **ReCoRD** (Zhang et al., 2018), given a passage  $p$  and a cloze question  $q$ , the task is to decide which

Model	Params (M)	BoolQ	CB	COPA	RTE	WiC	WSC	MultiRC	ReCoRD	Avg
		Acc.	Acc. / F1	Acc.	Acc.	Acc.	Acc.	EM / F1a	Acc. / F1	-
dev	GPT-3 Small	125	43.1	42.9 / 26.1	67.0	52.3	49.8	58.7	6.1 / 45.0	69.8 / 70.7
	GPT-3 Med	350	60.6	58.9 / 40.4	64.0	48.4	55.0	60.6	11.8 / 55.9	77.2 / 77.9
	GPT-3 Large	760	62.0	53.6 / 32.6	72.0	46.9	53.0	54.8	16.8 / 64.2	81.3 / 82.1
	GPT-3 XL	1,300	64.1	69.6 / 48.3	77.0	50.9	53.0	49.0	20.8 / 65.4	83.1 / 84.0
	GPT-3 2.7B	2,700	70.3	67.9 / 45.7	83.0	56.3	51.6	62.5	24.7 / 69.5	86.6 / 87.5
	GPT-3 6.7B	6,700	70.0	60.7 / 44.6	83.0	49.5	53.1	67.3	23.8 / 66.4	87.9 / 88.8
	GPT-3 13B	13,000	70.2	66.1 / 46.0	86.0	60.6	51.1	75.0	25.0 / 69.3	88.9 / 89.8
	GPT-3	175,000	77.5	82.1 / 57.2	92.0	72.9	<b>55.3</b>	75.0	32.5 / 74.8	<b>89.0 / 90.1</b>
test	PET	223	79.4	85.1 / 59.4	<b>95.0</b>	69.8	52.4	<b>80.1</b>	<b>37.9 / 77.3</b>	86.0 / 86.5
	iPET	223	<b>80.6</b>	<b>92.9 / 92.4</b>	<b>95.0</b>	<b>74.0</b>	52.2	<b>80.1</b>	33.0 / 74.0	86.0 / 86.5
	SotA	11,000	91.2	93.9 / 96.8	94.8	92.5	76.9	93.8	88.1 / 63.3	94.1 / 93.4
										89.3

Table 1: Results on SuperGLUE for GPT-3 primed with 32 randomly selected examples and for PET / iPET with ALBERT-xxlarge-v2 after training on FewGLUE. State-of-the-art results when using the regular, full size training sets for all tasks (Raffel et al., 2019) are shown in italics.

of a given set of answer candidates is the correct replacement for the placeholder in the cloze question. As this task is already presented in the form of a cloze question, there is little room for designing PVPs, so we only use a trivial one: the concatenation of  $p$  and  $q$  as pattern and the identity function as verbalizer. With only one PVP, there is no need to perform knowledge distillation so we directly use the resulting model as our final classifier.

## 4.2 Setup

As underlying LM for PET we choose ALBERT-xxlarge-v2 (Lan et al., 2020), the best-performing MLM on SuperGLUE when training is performed on the regular, full size training sets.<sup>5</sup> We run PET on the FewGLUE training sets for all SuperGLUE tasks using the exact same setup as Schick and Schütze (2020a). For COPA, WSC and ReCoRD, we use our proposed modification of PET to support verbalizers mapping labels to multiple tokens; for all other tasks, we use regular PET. We train iPET on all tasks except COPA and WSC, as their unlabeled sets contain well below 1000 examples, as well as ReCoRD, for which iPET makes no sense as we only use a single PVP. For these three tasks, we simply reuse the results of regular PET.

## 4.3 Results

Our main results are shown in Table 1. As can be seen, ALBERT with PET performs similar to the largest GPT-3 model, which is larger by a factor of 785. On average, PET performs 18 points better

compared to GPT-3 Med, a model of similar size. iPET brings further improvements for 3 out of 5 tasks, most notably for CB, but results in a slight performance drop for MultiRC. Similar to GPT-3, we find that PET does not perform above random chance for WiC, which is difficult to reformulate as a language modeling task. ReCoRD is the only task for which GPT-3 consistently performs better than both PET and iPET. Despite PET’s strong performance, it still clearly performs worse than a state-of-the-art model trained on the regular, full size SuperGLUE training set.

## 5 Analysis

We investigate the importance of several factors for good few-shot performance: the choice of patterns and verbalizers, the usage of both unlabeled and labeled data, and properties of the underlying language model. We also look into our proposed modification for PET to work with multiple masks and compare it to various baselines. Finally, we measure how choosing different sets of training examples affects performance. Our analysis focuses on PET as GPT-3 is not publicly available.<sup>6</sup>

### 5.1 Patterns

One factor that has gained little attention in previous work is the way tasks are reformulated as cloze questions. These reformulations can be arbitrarily complex; for example, the pattern used by GPT-3 for WSC contains an introductory section

<sup>5</sup>Results for various models can be found at <https://super.gluebenchmark.com/leaderboard>.

<sup>6</sup>We requested access to OpenAI’s GPT-3 API, but as of this writing, it has not been granted.

<b>Model</b>	<b>CB</b> Acc. / F1	<b>RTE</b> Acc.	<b>MultiRC</b> EM / F1a	<b>Avg</b> –
PET ( $\mathbf{p}_{\text{ours}}$ )	<b>85.1</b> / 59.4	69.8	37.9 / 77.3	66.6
PET ( $\mathbf{p}_{\text{GPT-3}}$ )	83.3 / 58.1	71.8	25.4 / 68.3	63.1
PET ( $\mathbf{p}_{\text{comb}}$ )	84.5 / 59.0	<b>74.7</b>	39.1 / <b>77.7</b>	68.3
PET ( $\mathbf{p}_{\text{ours}}$ ) - dist	83.9 / <b>76.2</b>	66.4	38.9 / 76.2	68.0
PET ( $\mathbf{p}_{\text{comb}}$ ) - dist	83.9 / <b>76.2</b>	72.9	<b>39.6</b> / 76.6	<b>70.4</b>

Table 2: Results on selected tasks for various sets of PVPs for regular PET and for an ensemble of PET models with no knowledge distillation (“- dist”)

that spans almost 30 words; it is unclear if and how this formulation has been optimized. To investigate the importance of patterns and verbalizers, we compare three different sets of PVPs: our initial set of PVPs as defined in Section 4.1 (denoted  $\mathbf{p}_{\text{ours}}$ ), the single PVP used by GPT-3 ( $\mathbf{p}_{\text{GPT-3}}$ ), and the combination of both ( $\mathbf{p}_{\text{comb}}$ ).

We train ALBERT using PET with all three sets of patterns; results for selected SuperGLUE tasks are shown in Table 2 (top). As can be seen, the PVP used by GPT-3 outperforms our PVPs on RTE whereas our initial set of patterns performs much better on MultiRC. These large differences in performance highlight the importance of finding good ways to express tasks as cloze questions. As it is extremely difficult to ascertain which patterns perform well without trying them on a large set of examples, a key challenge for few-shot approaches is to be able to compensate for PVPs that the used LM fails to understand well. As seen in the performance of the model trained with  $\mathbf{p}_{\text{comb}}$ , PET is able to do so: not only does combining all PVPs compensate for the worse performance of  $\mathbf{p}_{\text{ours}}$  on RTE and of  $\mathbf{p}_{\text{GPT-3}}$  on MultiRC, it even further improves average performance across the three tasks considered compared to the best-performing set of patterns. This clearly demonstrates the potential of carefully engineering a set of suitable patterns as opposed to just choosing a single formulation without means of evaluating its actual effectiveness.

## 5.2 Unlabeled Data Usage

Unlike GPT-3, PET requires unlabeled data to distill the knowledge of all models based on individual PVPs into a single classifier; for iPET, unlabeled data is additionally used to generate training sets for future generations. While the benefit of iPET is already analyzed in Table 1, we now investigate the importance of unlabeled data for regular PET. To this end, we compare the performance of the

<b>Model</b>	<b>CB</b> Acc. / F1	<b>RTE</b> Acc.	<b>MultiRC</b> EM / F1a	<b>Avg</b> –
PET	<b>85.1</b> / <b>59.4</b>	<b>69.8</b>	<b>37.9</b> / <b>77.3</b>	<b>66.6</b>
unsupervised	33.5 / 23.1	55.0	3.9 / 60.3	38.5
supervised	60.7 / 42.5	50.2	4.3 / 49.8	43.0
PET (XLNet)	<b>88.7</b> / <b>83.0</b>	<b>60.4</b>	<b>21.4</b> / <b>66.6</b>	<b>63.4</b>
Priming (XLNet)	56.3 / 37.7	49.5	– / –	–

Table 3: Results on selected tasks for various ways of using the labeled examples available in FewGLUE

final classifier in PET to that of directly using the ensemble of models corresponding to individual PVPs. While using this ensemble entirely removes the need for unlabeled data, the ensemble for  $k$  PVPs is larger than the distilled model by a factor of  $3 \cdot k$  as we follow the default setting of PET and train three models per PVP. However, even for a large number of PVPs the ensemble is smaller than GPT-3 by two orders of magnitude.

Results without distillation can be seen in Table 2 (bottom). Averaged across the three tasks, the ensemble performs even better than the distilled classifier. This shows that if the goal is only to achieve good performance, then unlabeled data is not necessary; however, it is required to obtain a single, lightweight model as final classifier.

Of course, there are further ways to leverage unlabeled data such as keeping an auxiliary language modeling objective during finetuning (Chronopoulou et al., 2019). While we leave investigating the impact of additionally using such methods to future work, we note that they can easily be applied to PET while there is no straightforward way to combine them with priming.

## 5.3 Labeled Data Usage

We next investigate the effect of how labeled data is used, which is one of the key differences between priming and PET. We first compare PET with regular supervised training (i.e., without using any patterns), and with a fully unsupervised model (Table 3). Given 32 examples, PET clearly outperforms both baselines, which is in line with findings by Schick and Schütze (2020a).

We next compare PET directly to priming. However, we cannot do so using ALBERT as it, like most pretrained MLMs, is only able to process sequences of up to 512 tokens, which is not enough for a set of 32 examples. As it theoretically supports sequences of arbitrary length, we instead use XLNet (Yang et al., 2019) for this compari-

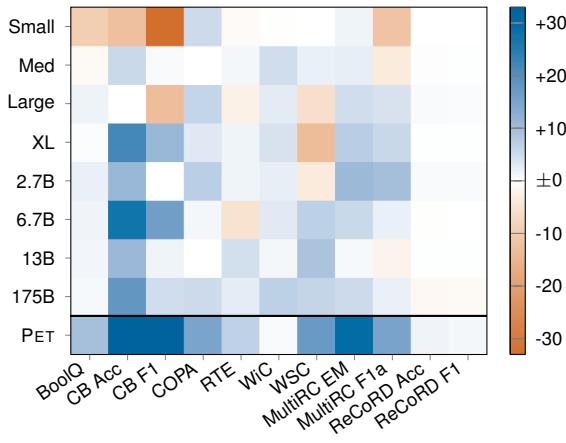


Figure 3: Accuracy differences between priming with 32 examples and one-shot priming for all GPT-3 models as well as between ALBERT with PET (without distillation) and unsupervised ALBERT (bottom row)

son. As shown in Table 3, XLNet in general performs worse than ALBERT. More importantly, XLNet with PET performs much better than priming (and priming does not even perform above random chance for RTE). We were not able to obtain results with priming on MultiRC because the 32 examples in FewGLUE would require more than 10,000 tokens, so processing them with a standard Transformer (Vaswani et al., 2017) is infeasible due to the quadratic complexity of self-attention. This highlights another important issue with priming: It does not scale well to more than a few examples; even GPT-3 is only able to process sequences of up to 2,048 tokens. While there are some variants of the Transformer that can deal with much longer contexts (e.g., Kitaev et al., 2020; Beltagy et al., 2020), it has yet to be investigated to what extent such models are able to actually make good use of priming examples over such a long context span.

We further investigate the effectiveness of priming by looking at results obtained with GPT-3 more closely. To analyze how well GPT-3 is able to make use of examples given as context, Figure 3 shows the performance difference between priming with 32 examples and priming with just a single example for each task and model size.<sup>7</sup> As can be seen, priming with 32 examples only slightly improves performance for most tasks and model sizes. For some tasks, adding more examples even leads to

<sup>7</sup> We do not compare priming to zero-shot performance as for unknown reasons, zero-shot GPT-3 performs well below random guessing for some tasks (e.g., 0.0% accuracy for WiC). To not overestimate the benefit of priming, we therefore show gains from providing 32 examples compared to just one.

Model	Params	CB		RTE	MultiRC	Avg
		Acc. / F1	Acc.	Acc.	EM / F1a	-
ALBERT	223M	<b>87.5 / 78.7</b>	<b>74.7</b>	<b>38.9 / 76.2</b>	<b>71.8</b>	
RoBERTa	355M	85.7 / 77.5	62.8	23.3 / 70.0	63.7	
GPT-2	345M	73.2 / 73.7	47.7	12.4 / 57.4	52.0	

Table 4: Results on selected tasks for PET without knowledge distillation combined with various LMs using  $\text{p}_{\text{GPT-3}}$  for CB/RTE and  $\text{p}_{\text{ours}}$  for MultiRC

worse performance, especially for smaller models. For ReCoRD, even the largest model’s performance slightly drops when adding more examples.

The bottom row of Figure 3 shows the performance difference between ALBERT trained with PET (without distillation) and a fully unsupervised ALBERT model on all tasks. While results are not directly comparable due to different underlying models and PVPs, using PET results in much stronger performance improvements compared to priming and does not worsen results for any of the eight tasks.

#### 5.4 Model Type

We next look into the relevance of the underlying LM for PET’s performance. To this end, we compare ALBERT with RoBERTa large (Liu et al., 2019) and GPT-2 medium (Radford et al., 2019). As GPT-2 is a unidirectional model similar to GPT-3, it can only process patterns where the mask token is the very last token. We therefore use  $\text{p}_{\text{GPT-3}}$  for CB and RTE; for MultiRC, we stick with our original set of patterns as they already fulfill this requirement. We also do not perform distillation and instead report the ensemble’s performance as there is no established way of equipping GPT-2 with a sequence classification head.

Results for training all three LMs with PET in Table 4 show that using ALBERT as underlying LM is crucial for PET’s strong performance; exchanging ALBERT with RoBERTa results in an average performance drop of 8 points. However, RoBERTa still clearly outperforms GPT-3 13B, which is larger by two orders of magnitude. Importantly, PET with GPT-2 performs much worse than with both other models. As anticipated by previous work (Brown et al., 2020), a key reason for this drop in performance could be that like GPT-3, GPT-2 is a unidirectional LM, making tasks that require comparing two sequences of text a much greater challenge. However, it is important to note that there are also other substantial differences between

<b>Model</b>	<b>COPA</b> Acc.	<b>WSC</b> Acc.	<b>ReCoRD</b> Acc. / F1	<b>Avg</b> –
PET	<b>95.0</b>	80.1	<b>86.0 / 86.5</b>	<b>87.1</b>
PET - dist (max-first)	90.0	<b>80.8</b>	<b>86.0 / 86.5</b>	85.7
PET - dist (ltr)	89.0	79.8	84.7 / 85.3	84.6
PET - dist (parallel)	77.0	<b>80.8</b>	82.5 / 83.1	80.2
untrained	72.5	59.9	84.7 / 85.4	72.5

Table 5: Results on selected tasks for our proposed variant of PET as well as other decoding strategies and for untrained ALBERT

GPT-2 and the other two models, most notably the dataset used for pretraining. Regardless of whether its unidirectionality is the reason for GPT-2’s bad performance, bidirectionality of the underlying LM is important for PET as it removes the need for the mask token to be at the very end and thus allows for more flexibility in the creation of patterns.

### 5.5 PET with Multiple Masks

We modified PET to work for outputs that require more than a single token. To investigate the impact of this modification, we look at the three tasks for which this is required: COPA, WSC and ReCoRD. We compare our decoding strategy of predicting tokens in order of the probability assigned to them, to which we refer as *max-first*, with two alternatives: decoding left-to-right (ltr) as is common for many autoregressive language models, and decoding all tokens simultaneously (parallel) as is done during training. Additionally, we compare PET with untrained ALBERT to measure the effectiveness of our proposed training loss.

Results are shown in Table 5. PET clearly outperforms untrained ALBERT for the three tasks. Not performing distillation hurts performance for COPA, but leads to slight improvements on WSC. Our decoding strategy is clearly superior to parallel decoding except for WSC, for which most predictions consist only of one or two tokens, and performs slightly better than left-to-right decoding.

### 5.6 Training Examples

Recall that we conduct our experiments with training examples from FewGLUE, a randomly selected subset of the original SuperGLUE training examples. We used a fixed random seed  $s_0$  to generate FewGLUE. Let  $\Sigma_i$  be the randomly selected subset of SuperGLUE for random seed  $s_i$ , so  $\Sigma_0 = \text{FewGLUE}$ . In this subsection, we create two additional subsets of SuperGLUE,  $\Sigma_1$  and  $\Sigma_2$ , based on different seeds. This allows us to investigate

<b>Model</b>	<b>CB</b> Acc. / F1	<b>RTE</b> Acc.	<b>MultiRC</b> EM / F1a	<b>Avg</b> –
GPT-3	82.1 / 57.2	<b>72.9</b>	32.5 / 74.8	65.4
PET - dist ( $\Sigma_0$ )	83.9 / 76.2	66.4	38.9 / 76.2	<b>68.0</b>
PET - dist ( $\Sigma_1$ )	82.1 / 57.4	61.4	<b>39.2 / 77.9</b>	63.2
PET - dist ( $\Sigma_2$ )	<b>87.5 / 84.0</b>	61.4	34.7 / 76.3	67.6

Table 6: Results on selected tasks for GPT-3 and for PET using training sets  $\Sigma_0$ ,  $\Sigma_1$ ,  $\Sigma_2$

how different sets of training examples affect performance. To this end, we run PET for CB, RTE and MultiRC using all three  $\Sigma_i$ . To measure only the effect of varying the training set while ignoring unlabeled examples, we do not use distillation.

Results in Table 6 show that for all tasks, changing the set of training examples can result in significant performance differences for PET. This highlights the importance of using the same set of examples when comparing different few-shot approaches, which is why we make the particular set of examples in FewGLUE publicly available. However, we note that the average performance of PET is similar to that of GPT-3 for all seeds.

While our results may seem contrary to the insight that for GPT-3, the exact choice of examples does not play a major role, we suspect this to be due to the fact that priming benefits much less from training examples than PET (cf. Section 5.3); accordingly, the influence of the exact set of training examples on the model’s performance is smaller.

## 6 Conclusion

We have shown that it is possible to achieve few-shot performance similar to GPT-3 on SuperGLUE with LMs that have three orders of magnitude fewer parameters. This is achieved using PET, a method that reformulates tasks as cloze questions and trains an ensemble of models for different reformulations. We have proposed a simple yet effective modification to PET that enables us to use it for tasks that require predicting multiple tokens. In extensive experiments, we have identified several factors responsible for the strong performance of PET combined with pretrained ALBERT: the possibility to concurrently use multiple patterns for transforming examples into cloze questions, the ability to compensate for patterns that are difficult to understand, the usage of labeled data to perform parameter updates, and the underlying LM itself. To enable comparisons with our work, we make our dataset of few-shot training examples publicly available.

## References

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. **Longformer: The long-document transformer.** *Computing Research Repository*, arXiv:2004.05150.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners.** *Computing Research Repository*, arXiv:2005.14165.
- Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. **An embarrassingly simple approach for transfer learning from pre-trained language models.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota. Association for Computational Linguistics.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. **BoolQ: Exploring the surprising difficulty of natural yes/no questions.** In *Proceedings of NAACL-HLT 2019*.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.
- Joe Davison, Joshua Feldman, and Alexander Rush. 2019. **Commonsense knowledge mining from pre-trained models.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ürün Dogan, Tobias Glasmachers, and Christian Igel. 2016. A unified view on multi-class support vector classification. *J. Mach. Learn. Res.*, 17(45):1–32.
- Allyson Ettinger. 2020. **What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models.** *Transactions of the Association for Computational Linguistics*, 8:3448.
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. **Mask-predict: Parallel decoding of conditional masked language models.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. **Distilling the knowledge in a neural network.** *Computing Research Repository*, arXiv:1503.02531.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. **How can we know what language models know?** *Computing Research Repository*, arXiv:1911.12543.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. **Reformer: The efficient transformer.** In *International Conference on Learning Representations*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. **ALBERT: A lite BERT for self-supervised learning of language representations.** In *International Conference on Learning Representations*.
- Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.

2019. RoBERTa: A robustly optimized BERT pre-training approach. *Computing Research Repository*, arXiv:1907.11692.
- Juri Opitz. 2019. Argumentative relation classification as plausibility ranking. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 193–202, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.
- Fabio Petroni, Tim Rocktischel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*.
- Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *Computing Research Repository*, arXiv:1912.10165.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Computing Research Repository*, arXiv:1910.10683.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2019. Learning semantic representations for novel words: Leveraging both form and context. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.
- Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676.
- Timo Schick and Hinrich Schütze. 2020b. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics – on what language model pre-training captures. *Computing Research Repository*, arXiv:1912.13283.
- Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *Computing Research Repository*, arXiv:1806.02847.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.
- Jason Weston and Chris Watkins. 1999. Support vector machines for multi-class pattern recognition. In *ESANN*, volume 99, pages 219–224.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art natural language processing. *Computing Research Repository*, arXiv:1910.03771.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc,

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *Computing Research Repository*, arXiv:1810.12885.

## A Training Details

Our implementation extends the original implementation of PET by Schick and Schütze (2020a) which, in turn, is based on the Transformers library (Wolf et al., 2019) and PyTorch (Paszke et al., 2017). Unless explicitly stated differently, we use the exact same set of hyperparameters as Schick and Schütze (2020a) with the only difference that for iPET, we only train 3 generations of models to speed up training. Below, we list task-specific implementation details for all tasks in SuperGLUE.

**COPA** For COPA, we randomly switch the two options  $c_1$  and  $c_2$  during training with a probability of 50% to make the input more diverse; for inference, we always keep the original order. For distilling the final PET model, we obtain logits for unlabeled examples  $x$  from individual PVPs  $\mathbf{p}$  as  $s_{\mathbf{p}}(y | x) = \log q_{\mathbf{p}}(y | x)$ ; we use the input format proposed by Liu et al. (2019).

**WiC** Similar to COPA, we randomly switch the input sentences  $s_1$  and  $s_2$  during training. Given a word  $w$  and two sentences  $s_1$  and  $s_2$ , we use the sequence  $w: s_1 | s_2$  as input for the final sequence classification model, where  $|$  marks the boundary between two text segments.

**WSC** Unlike other SuperGLUE tasks, the WSC formulation of Raffel et al. (2019) and Brown et al. (2020) requires free-form completion, meaning that for each sentence  $s$  and pronoun  $p$ , we only have a single correct choice  $n$  that the model needs to predict, but we do not provide any alternatives. During training, we thus use regular cross entropy loss between  $n$  and  $\tilde{q}_{\mathbf{p}}(n | s, p)$  as defined in Eq. 4. However, in many cases this would allow the LM to easily identify the correct target based on the number of masks provided, so we modify each target by randomly adding up to three additional mask tokens, for which we require the model to predict a special  $\langle \text{pad} \rangle$  token. For inference, we always just add a single mask token to ensure consistent results across multiple evaluations and perform greedy decoding as described in Section 3.

We then follow Raffel et al. (2019) to map the output produced by the LM to a label  $y \in \{\text{true}, \text{false}\}$ . For distillation, given an unlabeled example  $x$  we set  $s_{\mathbf{p}}(y | x) = 1$  if the model’s output for  $x$  was mapped to  $y$  and  $s_{\mathbf{p}}(y | x) = 0$  otherwise. We provide inputs to the final PET model in the format  $s | n$  where  $|$  is the boundary between two text segments and mark  $p$  in  $s$  with asterisks.

**MultiRC** Deviating from the hyperparameters used by Schick and Schütze (2019), we use a maximum sequence length of 512 tokens for MultiRC both during training and inference because we found many passages to be much longer than 256 tokens. Input for the final sequence classification model is of the form  $p | q | a$  where  $p$  is the passage,  $q$  is the question,  $a$  is the answer candidate and we use  $|$  to mark boundaries between text segments.

**ReCoRD** For ReCoRD, we again use a maximum sequence length of 512 because many passages require more than 256 tokens. For some questions  $q$ , the ReCoRD training set contains a huge number of answer candidates. To facilitate training, we split each example into multiple examples as follows: let  $C$  be the set of answer candidates with  $C^+ \subset C$  being the set of correct answers. We create a training example for each  $c \in C^+$  by randomly selecting up to 9 negative examples from  $C \setminus C^+$  for a total of 10 answer candidates.