# abstract

Prompts for pre-trained language models (PLMs) have shown remarkable performance by bridging the gap between pre-training tasks and various downstream tasks. Among these methods, prompt tuning, which freezes PLMs and only tunes soft prompts, provides an efficient and effective solution for adapting largescale PLMs to downstream tasks. However, prompt tuning is yet to be fully explored. In our pilot experiments, we find that prompt tuning performs comparably with conventional full-model fine-tuning when downstream data are sufficient, whereas it performs much worse under few-shot learning settings, which may hinder the application of prompt tuning in practice. We attribute this low performance to the manner of initializing soft prompts. Therefore, in this work, we propose to pretrain prompts by adding soft prompts into the pre-training stage to obtain a better initialization. We name this Pre-trained Prompt Tuning framework “PPT”. To ensure the generalization of PPT, we formulate similar classification tasks into a unified task form and pretrain soft prompts for this unified task. Extensive experiments show that tuning pre-trained prompts for downstream tasks can reach or even outperform full-model fine-tuning under both full-data and few-shot settings. Our approach is effective and efficient for using largescale PLMs in practice.

通过弥合预训练任务和各种下游任务之间的差距，预训练语言模型 (PLM) 的提示显示出卓越的性能。在这些方法中，冻结PLM并仅调整软提示的提示调整为使大规模PLM适应下游任务提供了高效且有效的解决方案。但是，提示调整尚未得到充分探索。在我们的试点实验中，我们发现当下游数据充足时，即时调整的性能与传统的全模型微调相当，而在少样本学习设置下它的表现要差得多，这可能会阻碍即时调整在实践中的应用。我们将这种低性能归因于初始化软提示的方式。因此，在这项工作中，我们建议通过在预训练阶段添加软提示来预训练提示以获得更好的初始化。我们将此预训练提示调优框架命名为“PPT”。为了保证PPT的泛化性，我们将相似的分类任务制定成统一的任务形式，并为这个统一的任务预训练软提示。大量实验表明，在全数据和少样本设置下，为下游任务调整预训练提示可以达到甚至超过全模型微调。我们的方法对于在实践中使用大规模 PLM 是有效和高效的。

# 1 introduction

Fine-tuning pre-trained language models (PLMs) (Devlin et al., 2019; Radford et al., 2019; Raffel et al., 2020) has made great progress in the recent years. By fine-tuning the entire parameters of PLMs, the versatile knowledge acquired from large-scale unlabeled corpora can be adapted to handle various NLP tasks and outperform the approach of learning models from scratch (Han et al., 2021a). For simplicity, we name this full-model tuning as “FT”. As shown in Figure 1 (b) and (c), there are two mainstream FT approaches. The first one is task-oriented fine-tuning, where a task-specific head is added on top of PLMs, and the entire model is then fine-tuned by optimizing task-specific learning objectives on task-specific training data.

微调预训练语言模型 (PLM)（Devlin 等人，2019 年；Radford 等人，2019 年；Raffel 等人，2020 年）近年来取得了很大进展。 通过微调 PLM 的整个参数，从大规模未标记语料库中获得的通用知识可以适应处理各种 NLP 任务，并优于从头开始学习模型的方法（Han 等人，2021a）。 为简单起见，我们将此全模型调整命名为“FT”。 如图1（b）和（c）所示，有两种主流的FT方法。 第一个是面向任务的微调，在 PLM 之上添加一个特定于任务的头，然后通过优化特定于任务的训练数据的特定于任务的学习目标来微调整个模型。

The second one is prompt-oriented finetuning (Schick and Schütze, 2021a), which is inspired by the recent works utilizing language prompts to stimulate the knowledge of PLMs (Petroni et al., 2019; Brown et al., 2020). In prompt-oriented fine-tuning, data samples are converted to linearized sequences containing prompt tokens, and all downstream tasks are formalized as language modeling problems. As shown in Figure 1 (c), by adding the prompt “It was hXi .” to a sentence, we can determine whether the sentence is positive or negative with PLMs predicting “great” or “terrible” at the mask position. As shown in Figure 1, compared to task-oriented fine-tuning, prompt-oriented fine-tuning is more similar to pretraining in terms of objectives (masked language modeling), thereby helping to better use knowledge in PLMs and often obtaining better performance.

第二个是面向提示的微调（Schick 和 Schütze，2021a），其灵感来自最近利用语言提示刺激 PLM 知识的作品（Petroni 等，2019；Brown 等，2020）。 在面向提示的微调中，将数据样本转换为包含提示标记的线性化序列，并将所有下游任务形式化为语言建模问题。 如图1(c)所示，通过添加提示“It was hXi”。 对于一个句子，我们可以通过 PLM 在掩码位置预测“很棒”或“可怕”来确定句子是肯定的还是否定的。 如图 1 所示，与面向任务的微调相比，面向提示的微调在目标（掩码语言建模）方面更类似于预训练，从而有助于更好地利用 PLM 中的知识并经常获得更好的性能。

Although the above-mentioned FT methods have shown promising results, with the rapid growth of model scale, fine-tuning a full large model for each downstream task becomes more and more expensive. To address this challenge, Lester et al. (2021) propose prompt tuning (PT) to adapt large PLMs to downstream tasks cheaply, as shown in Figure 1 (d). Specifically, PT uses soft prompts composed of continuous embeddings instead of hard prompts (discrete language phrases). These continuous prompt embeddings are generally randomly initialized and learned end-to-end. To avoid storing the entire model for each downstream task, PT freezes all parameters of PLMs and merely tune soft prompts, without adding any intermediate layers and task-specific components. Despite the few tunable parameters and the simple design, PT is competitive with FT, as illustrated in Figure 2(a)。

尽管上述 FT 方法已显示出可喜的结果，但随着模型规模的快速增长，为每个下游任务微调一个完整的大型模型变得越来越昂贵。 为了应对这一挑战，莱斯特等人。 (2021) 提出及时调整 (PT) 以低成本地使大型 PLM 适应下游任务，如图 1 (d) 所示。 具体来说，PT 使用由连续嵌入组成的软提示而不是硬提示（离散语言短语）。 这些连续提示嵌入通常是随机初始化和端到端学习的。 为了避免为每个下游任务存储整个模型，PT 冻结 PLM 的所有参数，仅调整软提示，而不添加任何中间层和特定于任务的组件。 尽管可调参数很少且设计简单，但 PT 与 FT 具有竞争力，如图 2(a) 所示

PT has two promising advantages: first, soft prompts can be learned end-to-end in comparison to hard prompts. Second, PT is an efficient and effective paradigm for the practical use of largescale PLMs. However, as shown in Figure 2(b), we find that PT performs much worse than FT under few-shot settings, which may hinder the application of PT in various low-resource scenarios.

PT 有两个有前途的优势：首先，与硬提示相比，软提示可以端到端地学习。 其次，PT 是大规模 PLM 实际使用的有效范式。 然而，如图 2(b) 所示，我们发现 PT 在少镜头设置下的表现比 FT 差很多，这可能会阻碍 PT 在各种低资源场景中的应用。

Hence, in this paper, we extensively explore how to use PLMs for few-shot learning in an efficient and effective manner through PT. More specifically, we conduct pilot experiments to empiri cally analyze the effectiveness of PT on large-scale PLMs for few-shot learning in Section 2, which is ignored by most existing works. Our discoveries are as follows: (1) the choice of verbalizer has a large impact on the performance; (2) simply initializing soft prompts with concrete word embeddings can not improve the performance, yet (3) combining soft and hard prompts is helpful; and (4) all these methods cannot handle few-shot prompt tuning problems well. The above observations reveal that finding suitable prompts for large-scale PLMs is not trivial, and carefully designed initialization of soft prompt tokens is crucial.

因此，在本文中，我们广泛探索了如何通过 PT 以高效和有效的方式使用 PLM 进行小样本学习。 更具体地说，我们在第 2 节中进行了试点实验，以实证分析 PT 在大规模 PLM 上的有效性，以进行小样本学习，大多数现有作品都忽略了这一点。 我们的发现如下：（1）verbalizer的选择对性能有很大的影响； (2) 简单地用具体的词嵌入初始化软提示并不能提高性能，但 (3) 软提示和硬提示的结合是有帮助的； (4) 所有这些方法都不能很好地处理少拍提示调整问题。 上述观察表明，为大规模 PLM 找到合适的提示并非易事，精心设计的软提示令牌初始化至关重要。

To help the model to find suitable prompts, we pre-train these tokens using self-supervised tasks on large-scale unlabeled corpora. To ensure the generalization of pre-trained prompts, we group typical classification tasks into three formats: sentence-pair classification, multiple-choice classification, and single-text classification, each format corresponding to one self-supervised pre-training task. In addition, we find multiple-choice classification is more general among these formats and we can unify all downstream classification tasks to this format. We name this Pre-trained Prompt Tuning (PPT) framework “PPT”. We evaluate PPT on several datasets using three 11B PLMs: T5-XXL (Raffel et al., 2020), mT5-XXL (Xue et al., 2021) and CPM-2 (Zhang et al., 2021b). Experiments show that PPT can not only improve few-shot PT by a large margin, reaching or even outperforming FT methods, but also reduce the variance of few-shot learning. Besides the effectiveness, PPT also retains the parameter efficiency of existing PT methods, which is valuable for future applications on large-scale PLMs.

为了帮助模型找到合适的提示，我们在大规模未标记语料库上使用自监督任务预训练这些标记。为了确保预训练提示的泛化，我们将典型的分类任务分为三种格式：句子对分类、多项选择分类和单文本分类，每种格式对应一个自监督的预训练任务。此外，我们发现多选分类在这些格式中更为普遍，我们可以将所有下游分类任务统一为这种格式。我们将此预训练提示调优 (PPT) 框架命名为“PPT”。我们使用三个 11B PLM 在几个数据集上评估 PPT：T5-XXL (Raffel et al., 2020)、mT5-XXL (Xue et al., 2021) 和 CPM-2 (Zhang et al., 2021b)。实验表明，PPT 不仅可以大幅度提高小样本 PT，达到甚至优于 FT 方法，而且可以减少小样本学习的方差。除了有效性之外，PPT 还保留了现有 PT 方法的参数效率，这对于未来在大规模 PLM 上的应用很有价值。

# 2 pilot experiments

In this section, we present several pilot experiments of PT under few-shot settings. We empirically analyze the effectiveness of three major categories of prompt enhancement strategies including hybrid prompt tuning, verbalizer selection, and real word initialization. We follow Lester et al. (2021) to test PT based on T5-XXL (with 11B parameters) and use 100 tunable soft-prompt tokens1 .

在本节中，我们将在少样本设置下展示 PT 的几个试点实验。 我们凭经验分析了三种主要类别的提示增强策略的有效性，包括混合提示调整、语言表达选择和实词初始化。 我们跟随莱斯特等人。 (2021) 测试基于 T5-XXL（具有 11B 参数）的 PT 并使用 100 个可调软提示标记1。

Following Schick and Schütze (2021a) and Schick and Schütze (2021b), we randomly select 32 samples to construct the training set Dtrain from the original training data and keep the samples across labels balanced. To tune the hyper-parameters and select the model, we compose a validation set Ddev from the original training data and ensure that |Dtrain| = |Ddev| to fit into a true few-shot learning setting (Perez et al., 2021). We follow Zhang et al. (2021a) and Gao et al. (2021) to use the original validation set as the test set Dtest, which means |Dtest| >> |Dtrain| = |Ddev|.

继 Schick and Schütze (2021a) 和 Schick and Schütze (2021b) 之后，我们随机选择 32 个样本从原始训练数据构建训练集 Dtrain，并保持样本跨标签平衡。 为了调整超参数并选择模型，我们从原始训练数据组成验证集 Ddev 并确保 |Dtrain| = |Ddev| 以适应真正的小样本学习环境（Perez 等人，2021 年）。 我们关注张等人。 (2021a) 和 Gao 等人。 (2021) 使用原始验证集作为测试集Dtest，即|Dtest| >> |Dtrain| = |Ddev|。

**Hybrid Prompt Tuning** In hybrid prompt tuning, both soft prompt tokens and hard prompt tokens are used (Liu et al., 2021; Han et al., 2021b). However, previous works train soft prompts together with the entire model. In the circumstances of PT, where only prompt tokens are tunable, the effectiveness of using hybrid prompts is underexplored. In Table 1, we show the results of combining soft prompt P with three manually designed hard prompts and two auto-generated hard prompts (Gao et al., 2021) on the sentiment classification task SST-2 (Socher et al., 2013). We can see that hard prompts improve PT, but still lag behind FT. Furthermore, different hard templates affect the performance a lot, for which much human labor for prompt design and selection is needed, providing a potential initialization for the next tuning.

混合提示调整在混合提示调整中，软提示标记和硬提示标记都被使用（Liu 等人，2021；Han 等人，2021b）。 然而，以前的工作与整个模型一起训练软提示。 在 PT 的情况下，只有提示令牌是可调的，使用混合提示的有效性尚未得到充分探索。 在表 1 中，我们展示了在情感分类任务 SST-2（Socher 等人，2013 年）上将软提示 P 与三个手动设计的硬提示和两个自动生成的硬提示（Gao 等人，2021 年）相结合的结果 . 我们可以看到硬提示提高了 PT，但仍然落后于 FT。 此外，不同的硬模板对性能影响很大，需要大量人工进行提示设计和选择，为下一次调优提供了潜在的初始化。

**Verbalizer Selection** How to choose the verbalizer that maps task-specific labels to concrete tokens is also worth studying. From Table 1 we can see that different choices of verbalizers influence the performance a lot. Generally, common words that explain the meaning of corresponding labels work well. This also guides our verbalizer selection for PPT in Section 3.

Verbalizer Selection 如何选择将特定任务标签映射到具体标记的语言器也值得研究。 从表 1 中我们可以看到，不同的语言表达器选择对性能影响很大。 通常，解释相应标签含义的常用词效果很好。 这也指导我们在第 3 节中为 PPT 选择语言表达工具。

Real Word Initialization The effectiveness of initializing soft prompts with the real word embeddings has been verified on small PLMs (fewer than 3B parameters) in previous works (Lester et al., 2021; Li and Liang, 2021). However, from the experiments on SST-2 (Socher et al., 2013) and a yes/no question answering task BoolQ (Clark et al., 2019) dataset (Table 2), we find that for the model with 11B parameters, real word initialization has little or even negative impact on the performance under few-shot settings. This suggests that observations on small models can not be directly transferred to large models and finding a good initialization for soft-prompt tokens is still crucial.

实词初始化 使用实词嵌入初始化软提示的有效性在之前的工作中已经在小型 PLM（少于 3B 参数）上得到验证（Lester 等人，2021 年；Li 和 Liang，2021 年）。 然而，从 SST-2（Socher 等人，2013 年）和是/否问答任务 BoolQ（克拉克等人，2019 年）数据集（表 2）的实验中，我们发现对于具有 11B 参数的模型， 实词初始化对少镜头设置下的性能影响很小甚至是负面影响。 这表明对小模型的观察不能直接转移到大模型，并且为软提示标记找到一个好的初始化仍然至关重要。

To summarize, although all the above three categories of prompt enhancement strategies cannot help PT achieve comparable results with FT under few-shot settings, the pilot experiments demonstrate the effectiveness of hybrid prompts, the good choice of the verbalizer, and the necessity of prompt initialization. In the following sections, we describe our PPT framework and show in experiments that PPT not only provides a good prompt initialization but also takes advantage of the good verbalizer and is complementary to hybrid prompts.

总而言之，虽然上述三类提示增强策略都不能帮助 PT 在少镜头设置下达到与 FT 相当的结果，但试点实验证明了混合提示的有效性、语言表达器的良好选择以及提示初始化的必要性 . 在接下来的部分中，我们将描述我们的 PPT 框架，并在实验中表明 PPT 不仅提供了良好的提示初始化，而且利用了良好的语言表达器，并且是对混合提示的补充。

# 3 pre-trained prompt tuning(PPT)

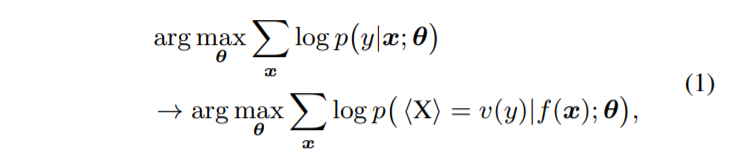
In this section, we describe the whole framework of PPT, including how to pre-train prompts and use these pre-trained prompts for specific tasks.

在本节中，我们描述了 PPT 的整个框架，包括如何预训练提示以及如何将这些预训练的提示用于特定任务。

## 3.1 Overview

Following the approach of T5 (Raffel et al., 2020) and PT (Lester et al., 2021), we solve all downstream tasks in a text-to-text format. As shown in Figure 1 (d), to reduce the objective gap between pre-training and downstream tasks, promptoriented fine-tuning converts downstream tasks into some cloze-style objectives. With a classification task as an example, given an input sentence x ∈ V∗ and its label y ∈ Y, a pattern mapping f : V ∗ 7→ V∗ is first applied to convert x into a new token sequence f(x), where V is the vocabulary of PLMs. f(x) not only adds some prompt tokens as hints, but also preserves at least one masking token hXi to let PLMs predict tokens at the masked positions. Then, a verbalizer v : Y 7→ V∗ is used to map y to a sequence of label tokens v(y). With f(·) and v(·), a classification task can be represented by a pattern-verbalizer pair (f, v):

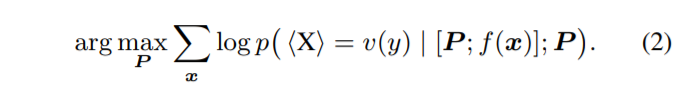
遵循 T5 (Raffel et al., 2020) 和 PT (Lester et al., 2021) 的方法，我们以文本到文本的格式解决所有下游任务。 如图 1(d) 所示，为了缩小预训练和下游任务之间的目标差距，prompt-oriented 微调将下游任务转换为一些完形填空式的目标。 以分类任务为例，给定输入句子 x ∈ V∗ 及其标签 y ∈ Y，首先应用模式映射 f : V ∗ 7→ V∗ 将 x 转换为新的标记序列 f(x)， 其中 V 是 PLM 的词汇表。 f(x) 不仅添加了一些提示标记作为提示，而且还保留了至少一个屏蔽标记 hXi 以让 PLM 预测屏蔽位置处的标记。 然后，使用语言表达器 v : Y 7→ V∗ 将 y 映射到标签标记序列 v(y)。 使用 f(·) 和 v(·)，分类任务可以用模式-语言表达器对 (f, v) 表示：



where θ indicates all tunable parameters, especially the parameters of PLMs. For convenience, we use “PVP” to denote this pattern-verbalizer pair (Schick and Schütze, 2021a).

其中 θ 表示所有可调参数，尤其是 PLM 的参数。 为方便起见，我们使用“PVP”来表示这个模式-言语者对（Schick 和 Schütze，2021a）。

In PT (Lester et al., 2021), a set of soft prompt tokens P are concatenated to the front of the sequence and the model input becomes [P ; f(x)], where [·; ·] is the concatenating function. By tuning P alone with other parameters fixed, Eq. (1) is replaced by  
在 PT (Lester et al., 2021) 中，一组软提示标记 P 连接到序列的前面，模型输入变为 [P ; f(x)]，其中 [·; ·] 是连接函数。 通过在固定其他参数的情况下单独调整 P，方程。 (1) 被替换为：



Owing to the power of large-scale PLMs, Eq. (2) is verified to be comparable to these FT methods under several full-data settings. However, we find that learning effective soft prompts is not easy, which may result in low performance under various fewshot settings. The parameter initialization usually has a large impact on the difficulty of learning models. Generally, besides randomly initializing p, some works sample word embeddings from the vocabulary of PLMs V as initialization. However, our pilot experiments have shown that existing initialization strategies and their simple variants have little or negative impact on the model performance based on large-scale PLMs. We refer more details of these pilot experiments to Section 4.

由于大规模 PLM 的强大功能，等式。 (2) 在几个全数据设置下被验证与这些 FT 方法相当。 然而，我们发现学习有效的软提示并不容易，这可能会导致在各种少拍设置下性能低下。 参数初始化通常对学习模型的难度有很大影响。 通常，除了随机初始化 p 之外，一些作品还从 PLM V 的词汇表中采样词嵌入作为初始化。 然而，我们的试点实验表明，现有的初始化策略及其简单的变体对基于大规模 PLM 的模型性能影响很小或产生负面影响。 我们将这些试点实验的更多细节参考第 4 节。

Recently, pre-training has been proven to be an effective method to find a good model initialization. Inspired by this, we propose to pre-train soft prompts. We notice that some groups of downstream tasks are related to certain self-supervised tasks built on unlabeled pre-training corpora. For instance, some tasks in the form of sentence-pair classification, such as natural language inference and sentence similarity, are similar to the next sentence prediction (NSP) (Devlin et al., 2019) task used in the pre-training stage. As shown in Figure 3, these tasks all take two sentences as input and compare their semantic meanings. Therefore, soft prompts pre-trained by NSP can be a good initialization for these sentence-pair tasks.

最近，预训练已被证明是找到一个好的模型初始化的有效方法。 受此启发，我们建议对软提示进行预训练。 我们注意到某些下游任务组与某些建立在未标记预训练语料库上的自监督任务相关。 例如，一些句子对分类形式的任务，如自然语言推理和句子相似性，类似于预训练阶段使用的下一句预测 (NSP) (Devlin et al., 2019) 任务。 如图3所示，这些任务都以两个句子作为输入，比较它们的语义。 因此，由 NSP 预训练的软提示可以很好地初始化这些句子对任务。

Formally, suppose we can divide downstream tasks into m groups {T1, T2, ..., Tm}, where Ti is the set containing ni downstream tasks: {PVP1 i ,PVP2 i , ..., PVPni i }, where PVPk i = (f k i , vk i ). For each group, we design one corresponding pre-training task PVPpre i = (f pre i , v pre i ). After pre-training soft prompts on these pretraining tasks with all model parameters fixed, we get m pre-trained prompts {P1, P2, ..., Pm}. After pre-training, for each task PVPk i in Ti , we continue to optimize Eq. (2) by using Pi as the initialization of soft prompts.

形式上，假设我们可以将下游任务分成 m 个组 {T1, T2, ..., Tm}，其中 Ti 是包含 ni 个下游任务的集合：{PVP1 i ,PVP2 i , ..., PVPni i }，其中 PVPk i = (fki , vk i )。 对于每一组，我们设计一个相应的预训练任务 PVPpre i = (f pre i , v pre i )。 在所有模型参数固定的这些预训练任务上预训练软提示后，我们得到 m 个预训练提示 {P1, P2, ..., Pm}。 预训练后，对于 Ti 中的每个任务 PVPk i，我们继续优化方程。 (2)通过使用Pi作为软提示的初始化。

## 3.2 Designing Pattern-Verbalizer Pairs for Pre-training

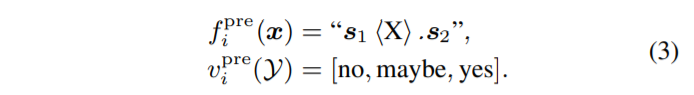
In this section, we take seveal typical classification tasks as an example to describe the design of pattern-verbalizer pairs PVPpre i for pre-training.

在本节中，我们以几个典型的分类任务为例来描述用于预训练的模式-语言对 PVPpre i 的设计。

### 3.2.1 Sentence-Pair Classification

Sentence-pair classification tasks such as natural language inference and sentence similarity take two sentences x = (s1, s2) as the input. To design a PVP for these tasks, we extend the next sentence prediction in Devlin et al. (2019) to a 3-class classification with labels Y = [0, 1, 2] as the pre-training task. These labels in Y can respectively indicate that the semantic relation between two sentences is coherent, similar and irrelevant. To construct signal from unlabeled pure text documents, we set the two sentences next to each other as label 2, those from the same document but not adjacent as 1, and those from different document as 0. We consider the label set |Y| <= 3 since this covers most sentence pair tasks. PVPpre i = (f pre i , v pre i ) is given as:

自然语言推理和句子相似度等句子对分类任务以两个句子 x = (s1, s2) 作为输入。 为了为这些任务设计 PVP，我们扩展了 Devlin 等人的下一句预测。 (2019) 将标签 Y = [0, 1, 2] 的 3 类分类作为预训练任务。 Y中的这些标签可以分别表示两个句子之间的语义关系是连贯的、相似的和不相关的。 为了从未标记的纯文本文档中构造信号，我们将相邻的两个句子设为标签 2，来自同一文档但不相邻的设为 1，来自不同文档的设为 0。我们考虑标签集 |Y| <= 3 因为这涵盖了大多数句子对任务。 PVPpre i = (f pre i , v pre i ) 给出如下：



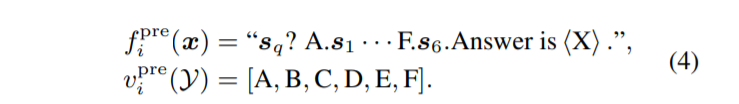
Designing PVPk i = (f k i , vk i ) according to PVPpre i is simple. s1 and s2 can be replaced by the input sentence pair. If a task outputs two labels, then we take v k i (Y) = [no, yes]. If a task outputs three labels, we set v k i = v pre i . If a task requires to measure the similarity between two sentences, the probability over {no, yes} can serve for this task.

根据 PVPpre i 设计 PVPk i = (f k i , vk i ) 很简单。 s1 和 s2 可以用输入句对代替。 如果一个任务输出两个标签，那么我们取 v k i (Y) = [no, yes]。 如果一个任务输出三个标签，我们设置 v k i = v pre i 。 如果一个任务需要测量两个句子之间的相似度，{no, yes} 的概率可以用于这个任务。

### 3.2.2 Multiple-Choice Classification

Many tasks can be formulated as the multiplechoice classification, which takes a query and several answer candidates as the input. We design a next sentence selection task to pre-train the prompt. Given a sentence as the query sq, the model is trained to select the adjacent sentence from six candidates, denoted as s1 ∼ s6 and thus the label set is Y = [1, 2, 3, 4, 5, 6]. These candidates consist of the right answer, one sentence from the same document but are not adjacent to the query, and four sentences from other documents. For x = (sq, s1, s2, · · · , s6), (f pre i , v pre i ) is given as

许多任务可以表述为多项选择分类，它以一个查询和几个候选答案作为输入。 我们设计了一个下一句选择任务来预训练提示。 给定一个句子作为查询 sq，模型被训练从六个候选中选择相邻的句子，表示为 s1 ∼ s6，因此标签集是 Y = [1, 2, 3, 4, 5, 6]。 这些候选包括正确答案、来自同一文档但不与查询相邻的一个句子以及来自其他文档的四个句子。 对于 x = (sq, s1, s2, · · · , s6), (f pre i , v pre i ) 给出为



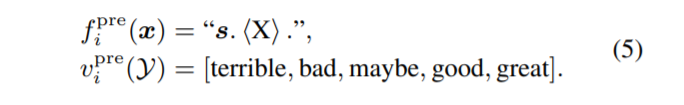
Most multiple-choice tasks can use {f pre i , v pre i } directly as their PVPs. For tasks like reading comprehension, the input may contain a passage and a question. We concatenate them to form a query.

大多数多项选择任务可以直接使用 {f pre i , v pre i } 作为它们的 PVP。 对于阅读理解等任务，输入可能包含一段话和一个问题。 我们将它们连接起来形成一个查询。

### 3.2.3 Single-Sentence Classification

For single-sentence classification, we create pseudo labels for prompt pre-training. Taking sentiment classification as an example, we use another small model to annotate sentiment labels for the sentences from the pre-training corpus and filter those with low classification probability. In practice, we use a RoBERTaBASE (Liu et al., 2019) model finetuned on a 5-class sentiment classification dataset other than the few-shot datasets we test on. Then with a sentence s from the corpus, we have the input x = (s) and the label set Y = [1, 2, 3, 4, 5]. (f pre i , v pre i ) is given as

对于单句分类，我们为提示预训练创建了伪标签。 以情感分类为例，我们使用另一个小模型对来自预训练语料库的句子进行情感标签标注，过滤掉分类概率较低的句子。 在实践中，我们使用 RoBERTaBASE (Liu et al., 2019) 模型在 5 类情感分类数据集上进行了微调，而不是我们测试的小样本数据集。 然后使用语料库中的句子 s，我们有输入 x = (s) 和标签集 Y = [1, 2, 3, 4, 5]。 (f pre i , v pre i ) 给出为



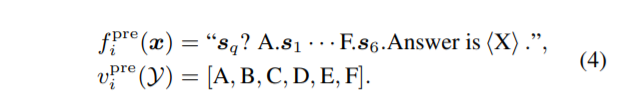
Designing PVPk i = (f k i , vk i ) according to PVPpre i is simple. s1 and s2 can be replaced by the input sentence pair. If a task outputs two labels, then we take v k i (Y) = [no, yes]. If a task outputs three labels, we set v k i = v pre i . If a task requires to measure the similarity between two sentences, the probability over {no, yes} can serve for this task.

根据 PVPpre i 设计 PVPk i = (f k i , vk i ) 很简单。 s1 和 s2 可以用输入句对代替。 如果一个任务输出两个标签，那么我们取 v k i (Y) = [no, yes]。 如果一个任务输出三个标签，我们设置 v k i = v pre i 。 如果一个任务需要测量两个句子之间的相似度，{no, yes} 的概率可以用于这个任务。

### 3.2.2 Multiple-Choice Classification

Many tasks can be formulated as the multiplechoice classification, which takes a query and several answer candidates as the input. We design a next sentence selection task to pre-train the prompt. Given a sentence as the query sq, the model is trained to select the adjacent sentence from six candidates, denoted as s1 ∼ s6 and thus the label set is Y = [1, 2, 3, 4, 5, 6]. These candidates consist of the right answer, one sentence from the same document but are not adjacent to the query, and four sentences from other documents. For x = (sq, s1, s2, · · · , s6), (f pre i , v pre i ) is given as

许多任务可以表述为多项选择分类，它以一个查询和几个候选答案作为输入。 我们设计了一个下一句选择任务来预训练提示。 给定一个句子作为查询 sq，模型被训练从六个候选中选择相邻的句子，表示为 s1 ∼ s6，因此标签集是 Y = [1, 2, 3, 4, 5, 6]。 这些候选包括正确答案、来自同一文档但不与查询相邻的一个句子以及来自其他文档的四个句子。 对于 x = (sq, s1, s2, · · · , s6), (f pre i , v pre i ) 给出为



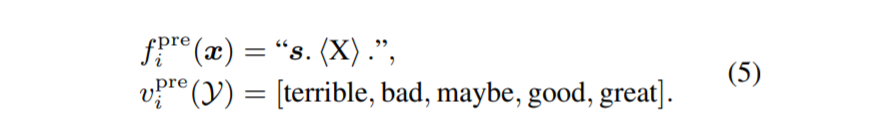
Most multiple-choice tasks can use {f pre i , v pre i } directly as their PVPs. For tasks like reading comprehension, the input may contain a passage and a question. We concatenate them to form a query.

大多数多项选择任务可以直接使用 {f pre i , v pre i } 作为它们的 PVP。 对于阅读理解等任务，输入可能包含一段话和一个问题。 我们将它们连接起来形成一个查询。

### 3.2.3 Single-Sentence Classification

For single-sentence classification, we create pseudo labels for prompt pre-training. Taking sentiment classification as an example, we use another small model to annotate sentiment labels for the sentences from the pre-training corpus and filter those with low classification probability. In practice, we use a RoBERTaBASE (Liu et al., 2019) model finetuned on a 5-class sentiment classification dataset other than the few-shot datasets we test on. Then with a sentence s from the corpus, we have the input x = (s) and the label set Y = [1, 2, 3, 4, 5]. (f pre i , v pre i ) is given as

对于单句分类，我们为提示预训练创建了伪标签。 以情感分类为例，我们使用另一个小模型对来自预训练语料库的句子进行情感标签标注，过滤掉分类概率较低的句子。 在实践中，我们使用 RoBERTaBASE (Liu et al., 2019) 模型在 5 类情感分类数据集上进行了微调，而不是我们测试的小样本数据集。 然后使用语料库中的句子 s，我们有输入 x = (s) 和标签集 Y = [1, 2, 3, 4, 5]。 (f pre i , v pre i ) 给出为



For sentiment classification tasks with 5 labels, we can use PVPk i = PVPpre i . For those tasks with fewer than 5 labels, we choose a subset from v pre i (Y) as labels. Although the above method improves the model performance, we have to point out that its generalization to other single-text classifications with different domains and numbers of labels is limited. However, the method described in the following section can effectively solve this problem.

对于有 5 个标签的情感分类任务，我们可以使用 PVPk i = PVPpre i 。 对于那些标签少于 5 个的任务，我们从 v pre i (Y) 中选择一个子集作为标签。 虽然上述方法提高了模型性能，但我们必须指出，它对其他具有不同领域和标签数量的单文本分类的泛化是有限的。 但是，下一节介绍的方法可以有效地解决这个问题。

## 3.3 Unifying Task Formats

The above-mentioned PVPs for pre-training can be unified to a single format: multiple-choice classification. Specifically, for the sentence-pair classification task, the query is the concatenation of the two sentences and there are three options: no, maybe, and yes. For single-sentence classification, the query is the input sentence and the options are the concrete labels. Note that in this way, the pre-trained PVPs can be used in single text classification tasks from arbitrary domains and with up to several labels.

上述用于预训练的 PVP 可以统一为一个格式：多选分类。 具体来说，对于句子对分类任务，查询是两个句子的连接，有三个选项：no、maybe 和 yes。 对于单句分类，查询是输入的句子，选项是具体的标签。 请注意，通过这种方式，预训练的 PVP 可以用于来自任意域和多达多个标签的单个文本分类任务。

Taking a unified PVP is similar to the idea of MultiQA (Talmor and Berant, 2019) and UnifiedQA (Khashabi et al., 2020). Recently, Zhong et al. (2021a) use some hard prompts to unify several tasks as a meta question answering task. They tune the entire model with this meta task on a collection of QA datasets and then transfer to other classification tasks in low-resource settings. However, our PPT focuses on only tuning soft prompts with the main body of PLMs fixed and our pretraining is conducted on fully unsupervised data, rather than the collection of supervised datasets.

采用统一的 PVP 类似于 MultiQA (Talmor and Berant, 2019) 和 UnifiedQA (Khashabi et al., 2020) 的思想。 最近，钟等人。 (2021a) 使用一些硬提示将多个任务统一为元问答任务。 他们在一组 QA 数据集上使用此元任务调整整个模型，然后转移到低资源设置中的其他分类任务。 然而，我们的 PPT 只专注于在 PLM 主体固定的情况下调整软提示，并且我们的预训练是在完全无监督的数据上进行的，而不是监督数据集的集合。

Since different tasks may have different candidate numbers and lengths, we construct pretraining samples with option numbers varying from 2 to 16 2 and option lengths from 50 to 20. We use the PVP in Section 3.2.2 for pre-training, and then apply pre-trained soft prompts to cover sentencepair classification, multiple-choice classification, and single-sentence classification.

由于不同的任务可能有不同的候选数量和长度，我们构建了选项数量从 2 到 16 2 和选项长度从 50 到 20 不等的预训练样本。我们使用 3.2.2 节中的 PVP 进行预训练，然后应用预训练 -经过训练的软提示涵盖句子对分类、多项选择分类和单句分类。

# 4 Experiments

In this section, we first describe our experimental setup to evaluate PPT. Then, we show the main results and analysis of our framework.

在本节中，我们首先描述我们的实验设置来评估 PPT。 然后，我们展示了我们框架的主要结果和分析。

## 4.1 Setup

We conduct experiments on both Chinese and English tasks (see Table 3). As described in Section 2, for tasks with fewer than 5 labels, we construct the training and validation set with 32 samples from the original training data and ensure the number of labels is balanced. For tasks with more than 5 labels like TNews and YahooAnswer, it is hard to compose a dataset with balanced samples across labels. Therefore, we randomly select 8 samples for each label.

For English datasets, we use T5-XXL with 11B parameters as our base model to do PT since previous work (Lester et al., 2021; Zhang et al., 2021b) have shown that, T5-XXL is comparable with FT in full-data setting. We also do FT experiments on various sizes of T5 to verify that T5-XXL performs better than other sizes in few-shot scenarios and improving prompt tuning based on T5-XXL is meaningful. For Chinese datasets, we do PT based on CPM-2. Since CPM-2 does not provide model with other sizes, we compare it with mT5 (Xue et al., 2021) of various sizes.

Consistently, we use 100 soft tokens for PT. As a result, the tunable parameters is only 100×4096 = 4.1 × 106 = 410K. Compared with the 11B (1.1 × 1010) parameters of FT, PT only needs to store 3000 times smaller parameters for each task.

我们对中文和英文任务进行了实验（见表 3）。如第 2 节所述，对于少于 5 个标签的任务，我们使用来自原始训练数据的 32 个样本构建训练和验证集，并确保标签数量平衡。对于 TNews 和 YahooAnswer 等超过 5 个标签的任务，很难用跨标签的平衡样本组合数据集。因此，我们为每个标签随机选择 8 个样本。

对于英语数据集，我们使用具有 11B 个参数的 T5-XXL 作为我们的基础模型来进行 PT，因为之前的工作（Lester 等人，2021 年；Zhang 等人，2021b）表明，T5-XXL 完全可以与 FT 相媲美- 数据设置。我们还对各种尺寸的 T5 进行了 FT 实验，以验证 T5-XXL 在少镜头场景中的表现优于其他尺寸，并且基于 T5-XXL 改进即时调整是有意义的。对于中文数据集，我们基于 CPM-2 进行 PT。由于 CPM-2 不提供其他尺寸的模型，我们将其与各种尺寸的 mT5 (Xue et al., 2021) 进行比较。

我们始终为 PT 使用 100 个软令牌。结果，可调参数只有 100×4096 = 4.1 × 106 = 410K。与 FT 的 11B（1.1×1010）参数相比，PT 只需要为每个任务存储 3000 倍小的参数。

## 4.2 Main Results

In this section, we present the main results of PPT. The results of English and Chinese datasets are shown in Table 4. In the row FT, we present the full-model fine-tuning results of the T5 model of various sizes. In the row PT, we show the results of PPT and other baselines. The first baseline is Vanilla PT, where the soft tokens are randomly initialized from a normal distribution. The second is the hybrid strategy in Section 2. We also consider LM Adaption used in Lester et al. (2021) in which the T5 model is further pre-trained for 10K steps with language modeling to reduce the gap between the pre-training and the fine-tuning. We also test two variants of PPT: Hybrid PPT, in which carefully designed hard prompts are combined with pre-trained soft prompt, and Unified PPT, in which all tasks are unified in the multiple-choice format.

在本节中，我们展示了 PPT 的主要结果。 英文和中文数据集的结果如表4所示。在FT行中，我们展示了各种尺寸的T5模型的全模型微调结果。 在 PT 行中，我们展示了 PPT 和其他基线的结果。 第一个基线是 Vanilla PT，其中软标记是从正态分布随机初始化的。 第二个是第 2 节中的混合策略。我们还考虑了 Lester 等人使用的 LM Adaption。 (2021) 其中 T5 模型通过语言建模进一步预训练 10K 步，以减少预训练和微调之间的差距。 我们还测试了 PPT 的两种变体：混合 PPT，其中精心设计的硬提示与预先训练的软提示相结合，以及统一 PPT，其中所有任务以多项选择格式统一。

Effectiveness From the Table 4 we have four observations. First, with the increase of the parameter number, the performance of FT improves. This means large-scale models still help in few-shot learning. Therefore, considering the intractable parameter number, we study PT on the large-scale pre-trained model. Note that for Chinese experiments, CPM-2 and mT5-XXL share the same parameter number. But CPM-2 outperforms mT5- XXL across all tasks. Therefore, we use CPM-2 as the base model.

有效性 从表 4 中，我们有四个观察结果。 首先，随着参数数量的增加，FT的性能有所提升。 这意味着大规模模型仍然有助于小样本学习。 因此，考虑到难以处理的参数数量，我们在大规模预训练模型上研究 PT。 请注意，对于中文实验，CPM-2 和 mT5-XXL 共享相同的参数编号。 但是 CPM-2 在所有任务中都优于 mT5-XXL。 因此，我们使用 CPM-2 作为基础模型。

Second, PPT outperforms Vanilla PT and LM Adaption across most datasets significantly. Although on BoolQ dataset, PPT lags behind Hybrid PT, simply combining PPT and hard template (Hybrid PPT) outperforms all baselines. This means pre-trained prompt and the idea of the hybrid prompt is complementary. Similar phenomenons also appear on other datasets like RACEm, LCQMC, and C3 , in which adding hard templates to PPT continues to improve results.

其次，PPT 在大多数数据集中明显优于 Vanilla PT 和 LM Adaption。 尽管在 BoolQ 数据集上，PPT 落后于 Hybrid PT，但简单地将 PPT 和硬模板（Hybrid PPT）结合起来就优于所有基线。 这意味着预先训练好的提示和混合提示的想法是互补的。 类似的现象也出现在 RACEm、LCQMC 和 C3 等其他数据集上，其中在 PPT 中添加硬模板继续提高结果。

Third, PPT outperforms FT for 10B models on all Chinese datasets and most English datasets. This indicates that there still remains a gap between masked language modeling and downstream tasks.

第三，PPT 在所有中文数据集和大多数英文数据集上的 10B 模型上优于 FT。 这表明掩码语言建模和下游任务之间仍然存在差距。

Pre-training soft prompt bridges this gap to some extend. Based on this observation, an intuitive extension of our method is to further pre-train the entire parameters using each PVPi pre and fine-tune the model to the corresponding downstream tasks. However, since we focus on prompt-tuning in this paper, we leave this idea to future work.

预训练软提示在一定程度上弥补了这一差距。 基于这一观察，我们方法的直观扩展是使用每个 PVPi 预训练进一步预训练整个参数，并将模型微调到相应的下游任务。 然而，由于我们在本文中专注于即时调整，我们将这个想法留给未来的工作。

Fourth, PPT results in lower variances on most of the datasets. Few-shot learning is notorious for its instability with becomes very obvious in Vanilla PT. For some datasets like SST-2, the variance reaches 15.5 which means model does not perform better than random guesses under some random seeds. Combining with hard prompt or further pretraining with language modeling can alleviate this problem to some extent. But on some datasets like CCPM, Hybrid PT increases the variance and LM Adaption does not guarantee the average performance. With the help of pre-training, the variance remains at a low level across all datasets.

第四，PPT 在大多数数据集上导致较低的方差。 小样本学习因其不稳定性而臭名昭著，在 Vanilla PT 中变得非常明显。 对于像 SST-2 这样的一些数据集，方差达到 15.5，这意味着模型在一些随机种子下的表现并不比随机猜测更好。 结合硬提示或进一步预训练与语言建模可以在一定程度上缓解这个问题。 但是在 CCPM 等一些数据集上，Hybrid PT 增加了方差，而 LM Adaption 并不能保证平均性能。 在预训练的帮助下，所有数据集的方差都保持在较低水平。

Unified PPT Unifying all formats to multiplechoice format is another variant of PPT. In Table 4, we can see that Unified PPT reaches comparable performance as PPT and Hybrid PPT, still outperforming soft-prompt tuning baselines. However, all the datasets we have considered so far have fewer than 5 classification labels. For tasks with more labels, especially single-text classification in which pseudo label pre-training is also not appropriate for cross-domain adaption, Unified PPT can be a good alternative. In Table 5, we test Unified PPT on datasets with more than 5 labels. For PT and FT, we use a verbalizer to map each label to its corresponding name. PT (MC) means we solve the task in a multiple-choice format without pre-training the prompt. We do not use the PPT for single-sentence classification in Section 3.2.3 because it is hard to find other suitable datasets to train the pseudo label annotator. However, we can see that Unified PPT still achieves the best performance, even exceeding FT by a large margin.

统一 PPT 将所有格式统一为多项选择格式是 PPT 的另一种变体。在表 4 中，我们可以看到统一 PPT 达到了与 PPT 和混合 PPT 相当的性能，仍然优于软提示调整基准。但是，到目前为止我们考虑的所有数据集的分类标签都少于 5 个。对于标签较多的任务，尤其是单文本分类，其中伪标签预训练也不适合跨域适应，Unified PPT 是一个不错的选择。在表 5 中，我们在超过 5 个标签的数据集上测试了统一 PPT。对于 PT 和 FT，我们使用动词化器将每个标签映射到其对应的名称。 PT (MC) 意味着我们以多项选择格式解决任务，而无需预先训练提示。我们在 3.2.3 节中没有使用 PPT 进行单句分类，因为很难找到其他合适的数据集来训练伪标签注释器。但是，我们可以看到Unified PPT仍然达到了最好的性能，甚至大大超过了FT。

4.3 Sample Efficiency

We discuss how FT, PT, and PPT compare when the number of training samples increases. In Figure 4, we show the trend of these methods on the RACEm and CB datasets. We can see that for 32 to 128 samples, PPT is consistently better than Vanilla PT, and the performances of the three methods gradually converge when the number grows to 256.

我们讨论了当训练样本数量增加时 FT、PT 和 PPT 的比较。 在图 4 中，我们展示了这些方法在 RACEm 和 CB 数据集上的趋势。 我们可以看到，对于 32 到 128 个样本，PPT 始终优于 Vanilla PT，并且当数量增长到 256 时，三种方法的性能逐渐收敛。

# 6 conclusion

In this paper, we present PPT, a framework that improves prompt tuning for few-shot learning. We propose to firstly unify downstream tasks to several formats. Then, we design self-supervised pretraining tasks for each format and pre-train the prompt on these tasks. Finally, we do prompt tuning on downstream tasks based on the initialization of the corresponding pre-trained prompts. Extensive experiments show that our method significantly outperforms other prompt tuning baselines, performing comparable or even better than fullmodel tuning.

在本文中，我们介绍了 PPT，这是一个改进小样本学习的快速调整的框架。 我们建议首先将下游任务统一为多种格式。 然后，我们为每种格式设计自我监督的预训练任务，并对这些任务的提示进行预训练。 最后，我们根据相应预训练提示的初始化对下游任务进行提示调整。 大量实验表明，我们的方法明显优于其他即时调整基线，性能与全模型调整相当甚至更好。

There are two important directions for future work: (1) Designing unified task formats and the corresponding pre-training objectives for other kind of tasks such as language generation and relation extraction. (2) Beyond the soft prompt, whether unified task pre-training helps the pretrained language models itself.

未来的工作有两个重要方向：（1）为其他类型的任务（如语言生成和关系提取）设计统一的任务格式和相应的预训练目标。 (2) 除软提示外，统一任务预训练是否有助于预训练语言模型本身。