Table 4: Details of the pretrained language models considered in this study. MLM, NSP, SOP, and NTP stand for masked language modeling, next sentence prediction, sentence order prediction, and next token prediction, respectively. It should be noted that ALBERT employs weight sharing, and its memory consumption is similar to BERT and RoBERTa.

Model		#Parameters	d_{hidden}	$d_{\text{embedding}}$	Pretraining task	Pretraining data
BERT	base	110M	768	768	MLM & NSP	BookCorpus, English Wikipedia
	large	340M	1024	1024		
RoBERTa	base	125M	768	768	MLM	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories
	large	355M	1024	1024		
ALBERT	base	12M	768	128	MLM & SOP	BookCorpus, English Wikipedia
	large	18M	1024	128		
GPT-2	small	117M	768	768	NTP	WebText
	base	345M	1024	1024		
	large	774M	1280	1280		

A Implementation Details

A.1 Details of the Models

Table 4 provides an overview of the language models used in this study, including base and large variants of BERT, RoBERTa, and ALBERT. Each model is trained with distinct pretraining tasks and datasets. In this study, we focus on transferring continuous prompts between masked language models, as this fill-in-the-blank mechanism is a natural way to probe knowledge (Shin et al., 2020). We also provide a preliminary empirical investigation of transferring continuous prompts between different model structures, e.g., from the encoder-only BERT model to the decoder-only GPT-2 model, which is discussed in §B.1.

Due to the variations in pretraining datasets and tokenizing methods, the language models in different families (e.g., BERT vs. RoBERTa) have different vocabularies. We obtained a shared vocabulary of tokens by taking the intersection of these individual vocabularies. During the transfer, we first encode the source prompt embeddings to the entire relative space. Then, we pick top-k dimensions of highest values ($k = 8192$) and set the rest of zero, which follows Norelli et al. (2022).

A.2 Details of the Projector Baseline

One of our baselines is a projector that maps the source embedding space to the target one. We trained a two-layer neural network as the projector based on the shared vocabulary. Specifically, we have

$$\text{Proj}(\boldsymbol{e}_i^{\text{s}}) = W_2(f(W_1\boldsymbol{e}_i^{\text{s}} + \boldsymbol{b}_1)) + \boldsymbol{b}_2, \tag{8}$$

where f is the Leaky ReLU activation function (Xu et al., 2015). For some anchor word i, we denote by $\boldsymbol{e}_i^{\text{s}}$ and $\boldsymbol{e}_i^{\text{t}}$ the word embeddings of the source model and target model, respectively. We train the projector by minimizing the mean squared error loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{k}\sum_{i=1}^{k}(\text{Proj}(\boldsymbol{e}_i^{\text{s}}) - \boldsymbol{e}_i^{\text{t}}), \tag{9}$$

where k is the size of shared vocabulary between two language models. We trained the neural network with 10 epochs using the Adam optimizer (Kingma & Ba, 2014). The learning rate was 5e-3 and the batch size was 16. The hidden dimension of this two-layer neural network was 768. We ran the validation on target models after each training epoch with the projected target prompt embeddings. We chose the projector with the highest validation performance and used it for test.

B Additional Results

In this appendix, we report preliminary results of the additional experiments conducted during the author response phase based on the reviewers' suggestions. In particular, we show the adaptability

Table 5: Results on transferring prompts between encoder and decoder models.

Method \ Target		$BERT_{base}$	$RoBERTa_{base}$	$GPT2_{small}$	$GPT2_{medium}$	$GPT2_{large}$
Direct tuning		50.56	46.24	31.62	32.23	34.44
Manual		30.64	20.48	4.73	8.01	10.23
Source	$BERT_{base}$	-	17.68	10.46	11.52	5.50
	$RoBERTa_{base}$	31.33	-	14.06	13.70	14.33
	$GPT2_{small}$	6.58	0.39	-	13.72	2.34
	$GPT2_{medium}$	4.06	0.50	5.02	-	1.79

Table 6: Results of transferring prompts from source models to $RoBERTa_{large}$ on the SST-2 and DPpedia classification tasks.

Method	SST-2 (accuracy)	DBpedia (accuracy)
Direct tuning	90.94	84.92
Manual	69.95	72.28
Source: $BERT_{base}$	82.45	77.05
Source: $RoBERTa_{base}$	84.63	80.81

of our method to different model architectures in §B.1, and experiment with classification tasks in §B.2.

B.1 Transfer Between Different Model Architectures

We first demonstrate the feasibility of transferring continuous prompts across different model architectures. This experiment explores the transferability between encoder and decoder models, focusing on generative GPT-2 models of varying sizes: small, medium, and large, as detailed in Table 4. We selected $BERT_{base}$ and $RoBERTa_{base}$, two encoder models, for our primary experiment to examine the transferability of prompts to or from GPT-2 models.

Table 5 shows the results of transferring continuous prompts across architectures on the LAMA dataset, including comparisons with the performance of directly tuned and manually prompted target models for reference. We see that the prompts induced on the encoder models, $BERT_{base}$ and $RoBERTa_{base}$, are transferable to the GPT-2 models with different sizes. Notably, $RoBERTa_{base}$ shows its best transferability, outperforming the manual prompting baseline across all target models. However, we found that the GPT-2 models as the source cannot induce as meaningful prompts as the encoder models, often underperforming manual prompting. The underlying reason contributing to the poor transferability of the continuous prompts induced on GPT-2 models remains unexplored and merits further study.

B.2 Results on Classification Tasks

Now we show our proposed transfer method is effective on other NLP tasks. Specifically, we include SST-2, a binary sentiment classification task, and DBpedia, a 14-category topic classification task. Unlike LAMA's entity prediction which requires the model to consider the whole vocabulary, the classification task only requires prediction within the label words based on the prompt, for example, "great" or "bad" for the SST-2 dataset (Sun et al., 2022).

As shown in Table 6, compared to using manual prompts on the target model directly, transferring prompts from both $BERT_{base}$ and $RoBERTa_{base}$ to the $RoBERTa_{large}$ target model yields better results. In line with our previous findings, $RoBERTa_{base}$ shows its superior transferability. Overall, our additional results present the potential of applying our approach to various tasks and model architectures.

ACKNOWLEDGMENTS

We would like to thank Ning Shi for his insightful suggestion of the multi-source transfer approach. We also thank all reviewers and chairs for their valuable and constructive comments. The research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN2020-04465, the Amii Fellow Program, the Canada CIFAR AI Chair Program, a UAHJIC project, an Alberta Innovates Program, and the Digital Research Alliance of Canada (alliancecan.ca).