A Survey on Prompt Tuning

Zongqian Li [1] **Yixuan Su** [1] **Nigel Collier** [1]

Abstract

This survey reviews prompt tuning, a parameter-efficient approach for adapting language models by prepending trainable continuous vectors while keeping the model frozen. We classify existing approaches into two categories: direct prompt learning and transfer learning. Direct prompt learning methods include: general optimization approaches, encoder-based methods, decomposition strategies, and mixture-of-experts frameworks. Transfer learning methods consist of: general transfer approaches, encoder-based methods, and decomposition strategies. For each method, we analyze method designs, innovations, insights, advantages, and disadvantages, with illustrative visualizations comparing different frameworks. We identify challenges in computational efficiency and training stability, and discuss future directions in improving training robustness and broadening application scope.[1]

1. Introduction

Large Language Models (LLMs) have achieved success across various Natural Language Processing (NLP) tasks, yet their adaptation to downstream applications faces challenges due to the computational and storage costs of full model fine-tuning (Zhao et al., 2024). Prompt tuning has emerged as a promising parameter-efficient fine-tuning (PEFT) approach that offers several advantages: (1) parameter efficiency through updating only a small group of continuous vectors while keeping the pretrained language model frozen; (2) modular adaptation through task-specific prompts that preserve the original model and parameters, enabling efficient deployment; (3) framework flexibility supporting various knowledge transfer and composition mechanisms, facilitating multi-task learning and domain adaptation. (Lester et al., 2021)

The growing significance of prompt tuning is evidenced by its dual influence: for the academic community, it intersects with various areas including model adaptation, knowledge transfer (Hinton et al., 2015), continual learning (Wang et al., 2024), and model interpretability (Wang et al., 2024); for commercial applications, it enables economical model customization, reduces deployment overhead, and accelerates application development. This widespread adoption across both domains necessitates a comprehensive survey to understand its underlying mechanisms, framework designs, innovations, and development trends.

Unlike existing surveys or papers that broadly review PEFT (Han et al., 2024) or general prompting methods (Schulhoff et al., 2024), this survey focuses on prompt tuning approaches and makes three primary contributions:

- We present a comprehensive categorization of prompt tuning approaches, categorizing existing methods into two main branches as shown in Figure 1: direct prompt learning and transfer learning. Direct prompt learning encompasses methods that perform single-stage training on target tasks. Transfer learning methods, on the other hand, focus on leveraging knowledge from source tasks to improve performance on target tasks.
- For each method, we provide an in-depth analysis of its framework, principles, innovations, key insights, and relative advantages and limitations;
- We provide illustrative visualizations to compare different methods.
- We identify challenges in current prompt tuning methods and project future directions.

This survey is organized as follows: Section 2 introduces the preliminaries of PEFT, soft prompts, and transfer learning. Section 3 and Section 4 present a review of direct prompt learning and transfer learning approaches, respectively. Section 5 discusses current challenges and future directions, followed by conclusions in Section 6.

2. Preliminary

PEFT aims to adapt pre-trained language models with minimal trainable parameters. These methods can be broadly

[1]University of Cambridge. Correspondence to: Zongqian Li, Nigel Collier <zl510@cam.ac.uk, nhc30@cam.ac.uk>.

3rd Workshop on Efficient Systems for Foundation Models at the 42nd International Conference on Machine Learning, Vancouver, Canada. Copyright 2025 by the author(s).

[1]https://github.com/ZongqianLi/Prompt-Compression-Survey

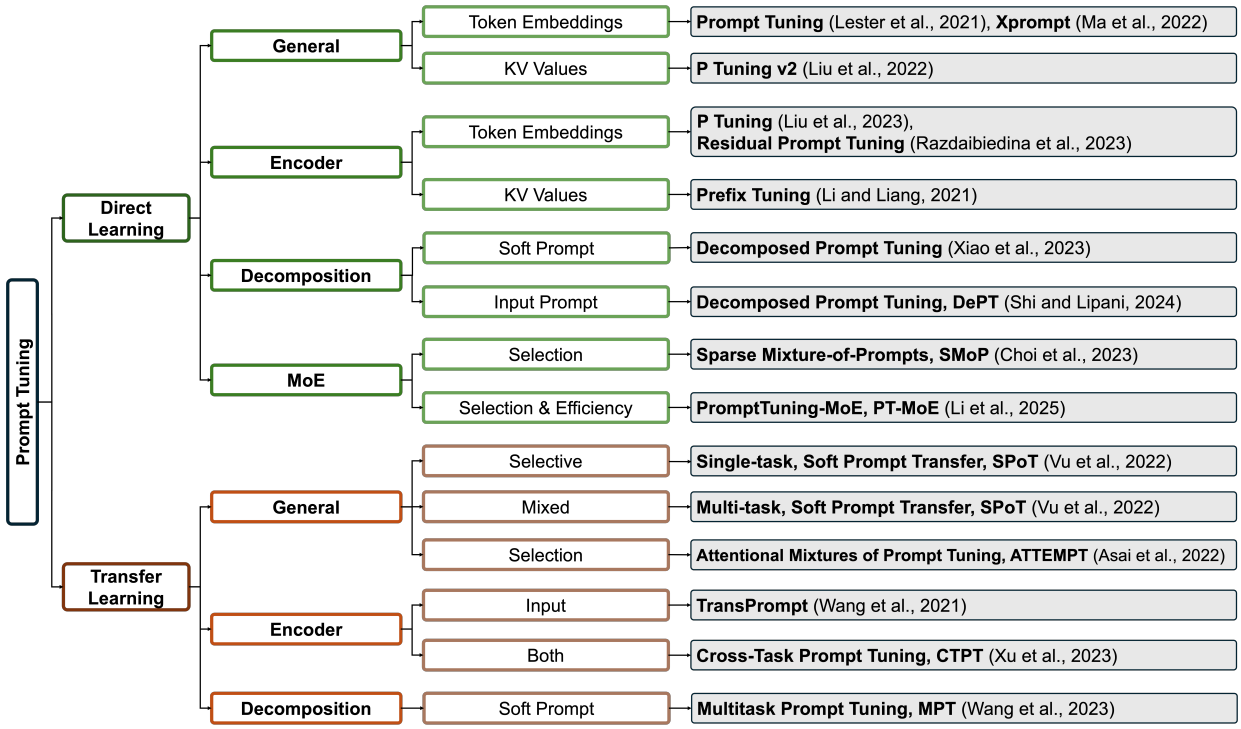

Figure 1: Hierarchical overview of prompt tuning methods including direct learning and transfer learning.

categorized into: (1) addition-based methods that introduce new parameters such as adapters inserted between transformer layers (Houlsby et al., 2019); (2) selection-based methods that update specific types of parameters such as bias terms while freezing other components (Ben Zaken et al., 2022); (3) reparameterization-based methods that decompose weight updates into low-rank matrices to reduce trainable parameters (Hu et al., 2022); (4) prompt-based methods that prepend trainable vectors to the input sequence (Lester et al., 2021). These approaches modify a small amount of model parameters while achieving comparable performance to full fine-tuning. Parameter-efficient methods reduce computational resources and memory requirements during training and deployment, making large-scale model adaptation more accessible (Wan et al., 2024). However, they often require careful hyperparameter tuning and may exhibit performance decrease on complex tasks or small-scale pre-trained models.

Soft prompts are continuous vectors in the embedding space, different from discrete prompts that consist of natural language tokens (with trained and fixed word embeddings) (Zhao et al., 2023). Formally, given an input sequence $\mathbf{x} = [x_1, ..., x_n]$, soft prompts $\mathbf{P} \in \mathbb{R}^{m \times d}$ are prepended to create a lengthened input sequence $[\mathbf{P}; E(\mathbf{x})]$, where $E(\cdot)$ denotes the embedding layer, m is the length of the soft prompt, and d is the dimension of the embedding space. Soft prompts have two key characteristics: (1) they are unrestricted by vocabulary constraints, enabling optimization in a continuous space and thus encoding task-specific information more flexibly than discrete prompts; (2) they maintain the same dimensionality as word embeddings, ensuring compatibility with transformer. These prompts can be initialized either randomly or from existing word embeddings, leading to different optimization trajectories in the embedding space.

Transfer learning in prompt tuning leverages knowledge from source tasks to improve performance on target tasks (Gu et al., 2022). Source tasks typically refer to large-scale datasets or tasks with abundant labeled data, while target tasks are usually low-resource scenarios where labeled data is limited (Zhuang et al., 2021). This framework comprises three essential components: (1) knowledge transfer strategies that address how to extract and encode knowledge from source tasks; (2) task adaptation mechanisms that determine how to utilize transferred knowledge for target tasks; (3) cross-domain methods that bridge distributional gaps between source and target domains, such as handling different label spaces. The effectiveness of transfer learning in prompt tuning depends on several variables: task similarity between source and target domains, quality of source task data, and the robustness of the transfer mechanism. Studies have proved that well-designed transfer approaches can improve prompt tuning performance, particularly in few-shot scenarios where target task data is limited.

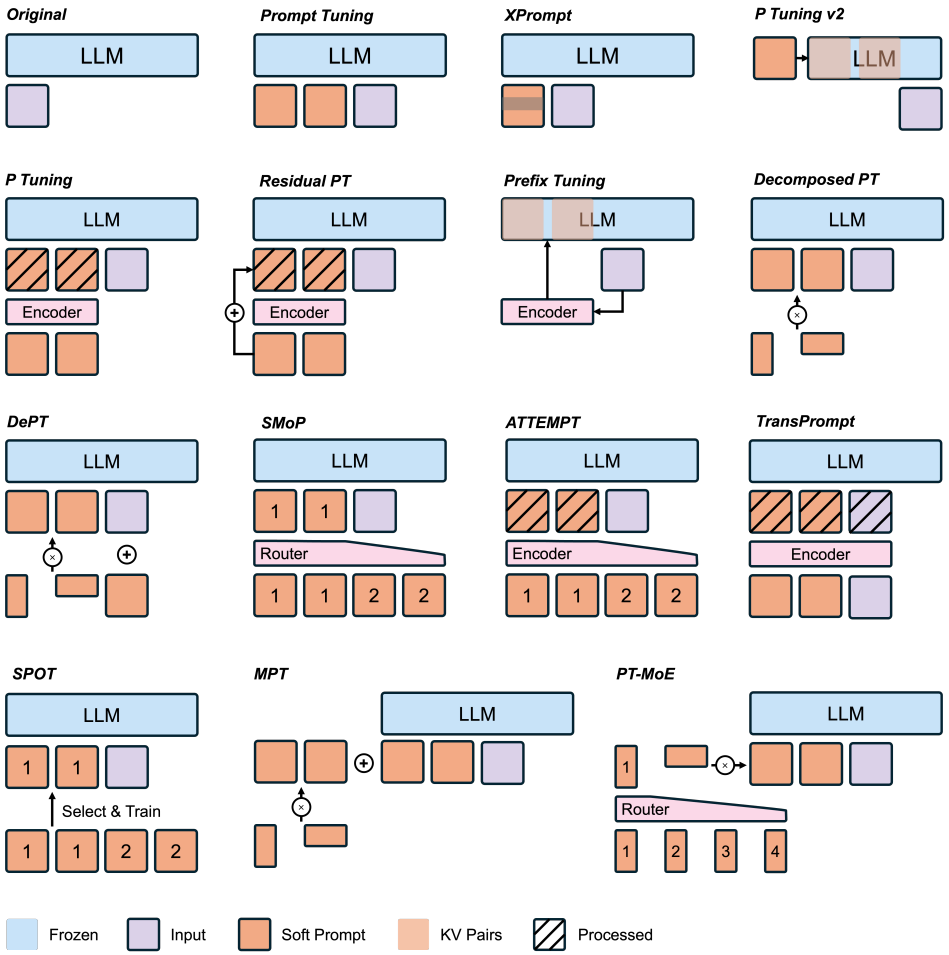

Figure 2: Illustration of different direct prompt tuning methods. Prompt Tuning directly prepends soft prompts to input. XPrompt applies pruning to soft prompts. P-Tuning v2 and Prefix Tuning incorporate prompts or KV pairs across all the layers of the language model. P-Tuning and Residual PT use encoders to process soft prompts. Decomposed PT and DePT leverage matrix decomposition strategies. SPoT leverages pre-trained prompts from source tasks, ATTEMPT or CTPT utilizes prompt mixing, TransPrompt uses encoders for task-specific and universal knowledge, while MPT decomposes prompts into shared and task-specific components.

3. Direct Prompt Learning

Direct prompt learning methods typically involve a single training stage where soft prompts are optimized on target tasks while keeping the pretrained language model parameters frozen. These approaches can be categorized into four main branches: General methods that directly tune prompt embeddings or KV pairs like Prompt Tuning (Lester et al., 2021), XPrompt (Ma et al., 2022), and P-Tuning v2 (Liu et al., 2022b); Encoder-based methods that incorporate additional layers to process prompts such as P-Tuning (Liu et al., 2023), Residual Prompt Tuning (RPT) (Razdaibiedina et al., 2023), and Prefix Tuning (Li & Liang, 2021); Decomposition-based methods that decompose prompt embeddings including Decomposed Prompt Tuning (DPT) (Xiao et al., 2023) and DePT (Shi & Lipani, 2024); and MoE approaches that leverage multiple soft prompts like Sparse Mixture-of-Prompts (SMoP) (Choi et al., 2023). Direct prompt learning methods are illustrated in Figure 2.

Prompt Tuning freezes all the parameters of the original pretrained language model and prepends a list of trainable tokens to the input tokens (Lester et al., 2021). These trainable tokens are initialized by random values or specific words embeddings, and are optimized to adapt the language model for downstream tasks. Prompt Tuning is the first method that finetunes the input tokens, which proves that training a small amount of parameters can reach similar performance with full finetuning. The larger the language model is, the more effective the method is. For T5-XXL with 11B parameters, prompt tuning matches the performance of full tuning within a couple percentage points on most tasks in

SuperGLUE. Less additional parameters save the time for training the language model and the memory to store the trained models. However, the training process for prompt tuning is difficult to converge and is sensitive to the initialization of the trainable tokens and learning rate. Sufficient training data are needed as well.

XPrompt is an efficient prompt tuning method that utilizes pruning to soft prompt tokens (Ma et al., 2022). This approach consists of three steps: (1) vanilla prompt tuning on the target tasks; (2) hierarchical pruning, which first removes negative soft prompt tokens at token-level and then uses fine-grained pruning at piece-level; (3) weight rewinding to retrain the identified positive soft prompt tokens. XPrompt is based on lottery ticket hypothesis. This theory proposes that an over-parameterized layer contains smaller sublayers that can achieve similar or better performance when trained independently from the same initialization. The key insight of XPrompt is the first application of lottery ticket hypothesis to prompt tuning, revealing that not all soft prompt tokens contribute equally to model performance and some tokens may even have negative influences. This observation leads to the innovation of piece-level pruning within each token. XPrompt improves model performance across different model scales, achieves higher parameter efficiency, and works well in both high-resource and few-shot scenarios. However, one question left is how to find the optimal compression ratio without trial training.

P-Tuning v2 improves prompt tuning by introducing deep prompts (soft prompts) across all layers of pre-trained language models while keeping the original parameters frozen (Liu et al., 2022b). Different from traditional prompt tuning that only adds trainable continuous prompts to the input layer, P-Tuning v2 incorporates prefix trainable tokens as soft prompts in every transformer layer. This deep prompt design eliminates the need for verbalizers and language model heads in favor of a simple classification head. The key insights of P-Tuning v2 include: (1) prompt depth influences model performance and adding soft prompts to deeper layers can match the performance of full-layer prompting; (2) optimal prompt length varies by task complexity with simple classification tasks preferring shorter soft prompts while sequence labeling tasks benefit from longer ones; (3) reparameterization influences are task-dependent rather than universally beneficial. P-Tuning v2 achieves comparable performance to full fine-tuning across different model scales (330M-10B) and diverse NLP tasks, demonstrating strong parameter efficiency. However, it requires task-specific hyperparameter tuning and may underperform full fine-tuning on certain sequence labeling tasks, particularly with smaller models.

P-Tuning introduces a hybrid method that uses trainable continuous prompt embeddings while incorporating discrete text tokens at specific positions, freezing the pretrained model parameters (Liu et al., 2023). Different from conventional prompt tuning, the continuous prompt embeddings in P-Tuning can be interrupted by discrete word tokens and a prompt encoder (LSTM or MLP) is employed to learn dependencies between continuous embeddings. There are several insights for P-Tuning: (1) continuous and discrete prompts complement each other, with continuous prompts providing optimization flexibility while discrete prompts maintain fiengrained information; (2) prompt placement influences task performance, with better results achieved when continuous prompts do not break the completeness of sentences; (3) the prompt encoder influences training stability with LSTM and MLP generally outperforming direct embedding optimization. P-tuning improves both model performance and training stability while reducing the variance between different prompts. However, this approach introduces additional trainable parameters for the prompt encoder, requires task-specific tuning of prompts, and shows decreasing returns on large-scale supervised tasks where traditional fine-tuning remains competitive.

RPT introduces a novel reparameterization approach that improves continuous prompt embeddings through a residual design while maintaining the frozen pretrained model (Razdaibiedina et al., 2023). The continuous embeddings are transformed through a residual layer before being prepended to inputs, where the MLP consists of down-projection and up-projection layers, followed by layer normalization and a skip connection to preserve the original embedding information. Several key insights emerge from this method: (1) the residual connection facilitates optimization by providing a direct path for preserving original embedding information while allowing the MLP to focus on learning useful transformations, leading to faster convergence and better performance; (2) the shared reparameterization layers exhibits superior performance in low-resource scenarios; (3) the compression strategy in the MLP design proves crucial for performance, with increased dimensionality leading to consistent improvements until saturation. RPT demonstrates improvements in performance stability and hyperparameter robustness. However, this approach still exhibits limitations: it introduces additional parameters through the reparameterization layer and maintains a performance gap with full fine-tuning.

Prefix-Tuning introduces a parameter-efficient approach that prepends trainable continuous key-value vectors at each transformer layer without modifying the pretrained model parameters (Li & Liang, 2021). Unlike prompt tuning that only optimizes embeddings at the input layer, Prefix-Tuning directly optimizes prefix keys and values across all transformer layers. Specifically, for each transformer layer, the prefix key-value pairs are generated through a reparametrization layer: a smaller trainable matrix is transformed through

an MLP to produce the layer-specific key-value vectors. Several key insights emerge from the empirical analysis: (1) direct optimization of prefix parameters leads to instability, necessitating the reparameterization strategy for robust training; (2) this method demonstrates superior performance in low-resource scenarios, surpassing full fine-tuning; (3) preserving pretrained parameters benefits cross-domain generalization. While Prefix-Tuning achieves parameter efficiency and enables efficient multi-task batching, it exhibits certain limitations: performance gap on large-scale datasets compared to full fine-tuning, and sensitivity to prefix length and initialization strategies.

DPT reparameterizes the soft prompts in prompt tuning through low-rank matrix decomposition (Xiao et al., 2023). Prompt tuning initializes soft prompts either randomly or from vocabulary embeddings, however, DPT decomposes the original prompt matrix $\mathbf{P}_{\text{emb}} \in \mathbb{R}^{e\times c}$ into the product of two smaller matrices $\mathbf{A} \in \mathbb{R}^{e\times b}$ and $\mathbf{B} \in \mathbb{R}^{b\times c}$, where b denotes a small bottleneck dimension. Several key insights emerge: (1) soft prompts naturally tend to converge towards low-rank forms during training, suggesting inherent dimensionality redundancy; (2) direct parameterization of this low-rank property through matrix decomposition offers a more effective way to learn prompt embeddings; (3) the bottleneck dimension influences the complexity of prompt embeddings. While DPT achieves parameter efficiency and performance improvements and demonstrates robust performance in few-shot scenarios, it exhibits certain limitations: inherited slow convergence from prompt tuning, sensitivity to bottleneck dimension selection, and unverified effectiveness on generation tasks beyond natural language understanding.

DePT improves the prompt tuning method through a decomposition strategy as well (Shi & Lipani, 2024). Unlike DPT which decomposes the prompt matrix through low-rank matrix decomposition, DePT proposes a hybrid approach that decomposes the original prompt matrix $\mathbf{P} \in \mathbb{R}^{l\times d}$ into a shorter soft prompt $\mathbf{P}_s \in \mathbb{R}^{m\times d}$ and a pair of low-rank matrices $\mathbf{A} \in \mathbb{R}^{s\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$, where r denotes the rank dimension and s is the maximum sequence length. The low-rank matrices are used to update the frozen word embeddings $\mathbf{W}$ through element-wise addition ($\mathbf{W}' = \mathbf{W} + \mathbf{BA}$), while the shorter soft prompt is prepended to the input sequence. The investigation reveals several key insights: (1) the combination of low-rank updates and soft prompts maintains competitive performance while enabling efficient parameter utilization due to reduced sequence length; (2) employing dual learning rates (α_1 for soft prompts and α_2 for low-rank matrices) is crucial for optimization convergence; (3) updating word embeddings through low-rank matrices improves the adaptability of inputs. DePT demonstrates advantages: it achieves performance improvements, reduces computational costs, shows increasing efficiency gains with larger models, and demonstrates robust integration with parameter-efficient transfer learning frameworks particularly in few-shot learning scenarios. However, it exhibits limitations: increased hyperparameter complexity due to dual learning rates, dependency on maximum sequence length for parameter allocation, and performance gaps compared to full fine-tuning in specific tasks.

SMoP reformulates the prompt tuning through mixture-of-prompts mechanism (Choi et al., 2023). While standard prompt tuning employs a single continuous soft prompt $\mathbf{P}_\theta \in \mathbb{R}^{l\times e}$ of length l, SMoP parameterizes multiple shorter prompts $\mathbf{P}_{\theta_j} \in \mathbb{R}^{l'\times e}$ where $l' \ll l$, coupled with a gating mechanism that dynamically routes each input embedding to the most suitable soft prompt. The investigation reveals several fundamental insights: (1) the performance gains from increasing prompt length become marginal beyond a large number of tokens, indicating inherent redundancy in long prompts; (2) the number of prompts k exhibits an optimal range determined by data availability, with excessive prompts leading to decreased performance due to insufficient training data per prompt; (3) maintaining balanced prompt utilization is crucial for optimal performance. While SMoP demonstrates efficiency improvements and achieves performance gains compared to standard prompt tuning, it faces limitations: the gating mechanism introduces additional parameters, and its effectiveness depends on sufficient training data for each prompt.

4. Transfer Learning

Transfer learning approaches in prompt tuning leverage knowledge from source tasks to improve performance on target tasks through prompt transfer or adaptation. These methods can be classified into three main groups: General methods that directly transfer or mix prompt knowledge like Soft Prompt Transfer (SPoT) (Vu et al., 2022) and Attentional Mixtures of Prompt Tuning (ATTEMPT) (Asai et al., 2022); Encoder-based methods that utilize additional layers to facilitate knowledge transfer such as TransPrompt (Wang et al., 2021) and Cross-Task Prompt Tuning (CTPT) (Xu et al., 2023); and Decomposition-based methods that decompose prompts into shared and task-specific components like Multitask Prompt Tuning (MPT) (Wang et al., 2023). Transfer learning methods for prompt tuning are illustrated in Figure 2.

SPOT introduces transfer learning into prompt tuning through two transfer strategies, both followed by continued training on target tasks (Vu et al., 2022). The first strategy, termed generic transfer, trains a general soft prompt on multiple source tasks and uses it to initialize prompts for all target tasks, which are then further tuned on the respective tasks. The second strategy, termed targeted transfer, adopts a retrieval-based approach: it first trains task-specific

prompts for each source task independently, preserving both early checkpoints (for task similarity comparison) and final checkpoints (for transfer initialization). For a new target task, it measures the similarity between this task and source tasks using the early checkpoints of the prompts for source tasks, selects the most similar source task, and initializes the target prompt using the final checkpoint of that task, followed by continued prompt tuning on the target task. Several key insights emerge from SPoT: (1) prompt transfer can effectively improve performance; (2) task similarity strongly correlates with transfer effectiveness, with high-quality source tasks being either large-scale datasets, complex reasoning tasks, or tasks similar to the target; (3) early checkpoints of prompts serve as better task embeddings than final checkpoints for measuring task similarity, suggesting the importance of capturing task-specific knowledge in the early training stage. While SPOT matches or outperforms full model fine-tuning across all model sizes while only tuning a small number of parameters, it exhibits certain limitations: high sensitivity to source task selection, incomplete capture of transfer characteristics through task embeddings, and additional computational overhead for computing task embeddings.

ATTEMPT introduces an attention-based prompt mixing mechanism for efficient multi-task knowledge transfer in prompt tuning (Asai et al., 2022). The method first pre-trains a group of source prompts $[\mathbf{P}_1, ..., \mathbf{P}_t]$ on large-scale source tasks. For each target task instance, ATTEMPT computes attention weights a_j between the input embedding and each prompt through an attention module G, which consists of down-projection, non-linear transformation, and up-projection layers. The instance-level prompt is then generated as a weighted combination: $\mathbf{P}_{\text{instance}} = \mathbf{P}_{\text{target}} + \sum_{j=1}^{t+1} a_j \mathbf{P}_j$, where $\mathbf{P}_{\text{target}}$ is a learnable target-task prompt. During training, only $\mathbf{P}_{\text{target}}$ and G are updated while keeping the language model frozen. Several key insights emerge from their investigation: (1) the effectiveness of prompt tuning exhibits strong scaling behavior and performance improves with larger base models despite poor results on smaller ones; (2) various source prompts contribute differently to target tasks, necessitating dynamic instance-level adaptation; (3) knowledge transfer exists between ostensibly different tasks, suggesting broader applicability of prompt reuse. While ATTEMPT achieves parameter efficiency and competitive performance, it faces certain limitations: increased memory overhead due to extended input sequences, degraded performance with small base models, and dependency on pre-trained source prompts.

TransPrompt reformulates prompt tuning as a cross-task knowledge transfer framework through a dual-encoder design (Wang et al., 2021). While normal prompt tuning methods may utilize a single prompt encoder, TransPrompt simultaneously employs two parallel encoders: a task-specific encoder and a universal encoder. Specifically, given an input x, both encoders process the same templated sequence. The task-specific encoder uses task-dependent prompt templates, while the universal encoder employs shared prompt templates. Both encoders are implemented as BiLSTM followed by MLPs. The outputs from both encoders are combined through average pooling before being fed into the language model for final prediction. Several key insights emerge from the investigation: (1) the dual-encoder design enables simultaneous capture of task-specific and universal knowledge; (2) debiasing mechanisms, specifically prototype-based and entropy-based techniques, are crucial for learning task-agnostic embeddings; (3) the universal prompt template effectively captures cross-task commonalities, facilitating generalization to unseen tasks. While TransPrompt demonstrates consistent performance improvements and exhibits strong generalization capabilities, it faces certain limitations: complex training procedure; computational overhead from the double encoder design, dependency on task similarity for effective transfer, and sensitivity to hyperparameters, particularly requiring careful learning rate tuning for optimal convergence.

CTPT introduces a cross-task prompt tuning framework comprising three key components: Task-specific Prompt Tuning (TSPT), Cross-Task Prompt Learning (CTPL), and Cross-Task Prompt Observation (CTPO) (Xu et al., 2023). The method leverages both task-specific prompts $\mathbf{P}_i$ trained on few-shot samples of the target task and a group of source prompts $\mathbf{P}_j$ from pre-trained related tasks. Specifically, TSPT learns task-specific knowledge independently, while CTPL captures cross-task knowledge through multi-head attention where $\mathbf{P}_i$ serves as the query while $\mathbf{P}_j$ provide keys and values, followed by a residual connection: $\mathbf{P}_{\text{final}} = \mathbf{P}_i + \sum_j \text{softmax}(\frac{\mathbf{W}_Q \mathbf{P}_i \mathbf{W}_K^T \mathbf{P}_j}{\sqrt{d}}) \mathbf{W}_V \mathbf{P}_j$. CTPO then filters redundant information through a gate-like mechanism. The attention parameters $\mathbf{W} = [\mathbf{W}_Q; \mathbf{W}_K; \mathbf{W}_V]$ are obtained through a dimension reduction strategy: mapping from a low-dimensional vector $\mathbf{z}$ to the parameter space via a fixed projector matrix $\mathbf{A}$, $\mathbf{W} = \mathbf{A}\mathbf{z}$. The entire framework is optimized through derivative-free optimization (CMA-ES) in low-dimensional space. Several key insights emerge from the investigation: (1) emotional knowledge is inherently transferable across conversation datasets despite varying label nomenclatures; (2) performance improvement exhibits a linear correlation with the number of source tasks, indicating scalable knowledge transfer; (3) the intrinsic dimensionality of prompt parameters is substantially lower than their apparent dimensions. While CTPT demonstrates computational efficiency and robust zero-shot transfer capabilities, it exhibits certain limitations: requiring multiple forward passes for optimization, slower convergence compared to gradient-based methods, and decreased performance when source and target tasks differ significantly.

MPT reformulates prompt tuning through a decomposition of soft prompts into shared and task-specific components via knowledge distillation (Wang et al., 2023). While conventional prompt tuning learns independent prompts for each task, MPT parameterizes the task prompt $\hat{\mathbf{P}}_k$ as: $\hat{\mathbf{P}}_k = \mathbf{P}^* \circ (\mathbf{u}_k \otimes \mathbf{v}_k^T)$ where $\mathbf{P}^* \in \mathbb{R}^{l \times d}$ is the shared prompt matrix, $\mathbf{u}_k \in \mathbb{R}^l$ and $\mathbf{v}_k \in \mathbb{R}^d$ form the rank-one task-specific matrix through outer product, $\circ$ denotes the Hadamard (element-wise) product, l denotes prompt length and d denotes embedding dimension. The framework employs a distillation objective combining KL-divergence for logit matching and MSE loss for hidden state to transfer knowledge from independently trained teacher prompts. During source training, MPT adopts a stochastic task sampling strategy that dynamically varies the number of tasks per batch. For target adaptation, it implements a dual learning rate scheme where task-shared and task-specific components are updated with distinct learning rates. Several key insights emerge from this investigation: (1) cross-task knowledge can be effectively compressed into a single shared prompt while preserving task-specific adaptability; (2) the decomposition of prompts into shared and task-specific components enables efficient knowledge transfer across tasks; (3) knowledge distillation from independently trained prompts facilitates learning better decomposable embeddings. While MPT demonstrates parameter efficiency and achieves competitive performance including few-shot scenarios, it exhibits certain limitations: complex training process, computational overhead from teacher model training and distillation process, sensitivity to prompt length selection, and suboptimal performance on complex tasks compared to full fine-tuning.

5. Challenges and Future Work

5.1. Current Challenges

Computational efficiency. The addition of soft prompts extends input sequence length, increasing memory consumption and computational costs. This overhead becomes substantial when processing long sequences or storing multiple task-specific prompts (Tay et al., 2022).

Training instability. Prompt tuning demonstrates high sensitivity to hyperparameter choices, particularly learning rates, leading to convergence difficulties and inconsistent performance across different initializations (Gu et al., 2022). This necessitates hyperparameter tuning to achieve stable and optimal results.

Prompt initialization. Initialization strategies for soft prompts, whether random or embedding-based, influence model performance and convergence, yet current approaches can yield suboptimal results and require tuning (Liu et al., 2022a).

Model scale dependency. The effectiveness of prompt tuning correlates with model scale, performing well on large models but decreasing on smaller ones, limiting its applicability in resource-constrained settings (Lester et al., 2021).

Explainability. The semantic meaning of learned prompts and their interaction mechanisms with pretrained models remain poorly understood, limiting their improvements.

5.2. Future Directions

Training stability. Future work should focus on developing more robust optimization strategies and adaptive learning techniques (Liu et al., 2020) that reduce hyperparameter sensitivity. This includes exploring adaptive prompt designs and innovative initialization strategies that ensure stable convergence.

Scope extension. Prompt tuning can extend to diverse-input tasks (Yao et al., 2024), multi-output prediction problems, and specific domains.

Transfer learning. Future work should explore meta-learning approaches (Nichol et al., 2018) for prompt adaptation, hierarchical prompt designs, and methods for handling different label spaces across tasks.

Interpretability. Key directions include analyzing optimization paths, developing explainability methods, and understanding connections with other PEFT approaches (Han et al., 2024).

Parameter efficiency. Key areas include developing lightweight prompt embeddings (Li et al., 2024b;a), exploring dynamic generation methods, and investigating shared prompt designs.

6. Conclusions

This survey has reviewed prompt tuning methods across direct learning and transfer learning. Through analysis of various framework innovations, we observe trends toward improved parameter efficiency, adaptive cross-task learning, and optimized training stability. While progress has been made, challenges remain in resource consumption, explainability, and model scale dependency. Future directions include robust optimization strategies, scope extension, and interpretability analysis.

Ethics Statement

This work does not involve ethical concerns.

Availability Statement

Papers and updates related to this survey can be found at https://github.com/ZongqianLi/Prompt-Tuning-Survey.

References

Asai, A., Salehi, M., Peters, M., and Hajishirzi, H. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6655–6672, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.446. URL https://aclanthology.org/2022.emnlp-main.446.

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.1. URL https://aclanthology.org/2022.acl-short.1.

Choi, J.-Y., Kim, J., Park, J.-H., Mok, W.-L., and Lee, S. SMoP: Towards efficient and effective prompt tuning with sparse mixture-of-prompts. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14306–14316, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.884. URL https://aclanthology.org/2023.emnlp-main.884.

Gu, Y., Han, X., Liu, Z., and Huang, M. PPT: Pre-trained prompt tuning for few-shot learning. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8410–8423, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.576. URL https://aclanthology.org/2022.acl-long.576.

Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lIsCS8b6zj.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Li, Z., Liu, Y., Su, Y., and Collier, N. Prompt compression for large language models: A survey, 2024a. URL https://arxiv.org/abs/2410.12388.

Li, Z., Su, Y., and Collier, N. 500xcompressor: Generalized prompt compression for large language models, 2024b. URL https://arxiv.org/abs/2408.03094.

Li, Z., Su, Y., and Collier, N. Pt-moe: An efficient fine-tuning framework for integrating mixture-of-experts into prompt tuning, 2025. URL https://arxiv.org/abs/2505.09519.

Liu, H., Tam, D., Mohammed, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=rBCvMG-JsPd.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate

and beyond. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgz2aEKDr.

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. Gpt understands, too, 2023. URL https://arxiv.org/abs/2103.10385.

Ma, F., Zhang, C., Ren, L., Wang, J., Wang, Q., Wu, W., Quan, X., and Song, D. XPrompt: Exploring the extreme of prompt tuning. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11033–11047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.758. URL https://aclanthology.org/2022.emnlp-main.758.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms, 2018. URL https://arxiv.org/abs/1803.02999.

Razdaibiedina, A., Mao, Y., Khabsa, M., Lewis, M., Hou, R., Ba, J., and Almahairi, A. Residual prompt tuning: improving prompt tuning with residual reparameterization. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6740–6757, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.421. URL https://aclanthology.org/2023.findings-acl.421.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., Costa, H. D., Gupta, S., Rogers, M. L., Goncearenco, I., Sarli, G., Galynker, I., Peskoff, D., Carpuat, M., White, J., Anadkat, S., Hoyle, A., and Resnik, P. The prompt report: A systematic survey of prompting techniques, 2024. URL https://arxiv.org/abs/2406.06608.

Shi, Z. and Lipani, A. DePT: Decomposed prompt tuning for parameter-efficient fine-tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KjegfPGRde.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), December 2022. ISSN 0360-0300. doi: 10.1145/3530811. URL https://doi.org/10.1145/3530811.

Vu, T., Lester, B., Constant, N., Al-Rfou', R., and Cer, D. SPoT: Better frozen model adaptation through soft prompt transfer. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5039–5059, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.346. URL https://aclanthology.org/2022.acl-long.346.

Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., Zhang, Q., Chowdhury, M., and Zhang, M. Efficient large language models: A survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bsCCJHbO8A. Survey Certification.

Wang, C., Wang, J., Qiu, M., Huang, J., and Gao, M. TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2792–2802, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.221. URL https://aclanthology.org/2021.emnlp-main.221.

Wang, L., Zhang, X., Su, H., and Zhu, J. A Comprehensive Survey of Continual Learning: Theory, Method and Application . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(08):5362–5383, August 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3367329. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2024.3367329.

Wang, Z., Panda, R., Karlinsky, L., Feris, R., Sun, H., and Kim, Y. Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Nk2pDtuhTq.

Xiao, Y., Xu, L., Li, J., Lu, W., and Li, X. Decomposed prompt tuning via low-rank reparameterization. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pp. 13335–13347, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.890. URL https://aclanthology.org/2023.findings-emnlp.890.

Xu, Y., Zeng, Z., and Shen, Z. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11654–11666, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.780. URL https://aclanthology.org/2023.findings-emnlp.780.

Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.-S., and Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2024.01.004. URL https://www.sciencedirect.com/science/article/pii/S2666651024000056.

Zhao, W., Gupta, A., Chung, T., and Huang, J. SPC: Soft prompt construction for cross domain generalization. In Can, B., Mozes, M., Cahyawijaya, S., Saphra, N., Kassner, N., Ravfogel, S., Ravichander, A., Zhao, C., Augenstein, I., Rogers, A., Cho, K., Grefenstette, E., and Voita, L. (eds.), *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pp. 118–130, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.repl4nlp-1.10. URL https://aclanthology.org/2023.repl4nlp-1.10.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

A. Appendix

List of papers:

Direct Prompt Learning:
General:
Prompt Tuning (Lester et al., 2021), XPrompt (Ma et al., 2022), P Tuning v2 (Liu et al., 2022b)
Encoder:
P Tuning (Liu et al., 2023), Residual Prompt Tuning (Razdaibiedina et al., 2023), Prefix Tuning (Li & Liang, 2021)
Decomposirion:
Decomposed Prompt Tuning (Xiao et al., 2023), DePT (Shi & Lipani, 2024)
MoE:
SMoP (Choi et al., 2023), PT-MoE (Li et al., 2025)
Transfer Learning:
General:
SPoT (Vu et al., 2022), ATTEMPT (Asai et al., 2022)
Encoder:
TransPrompt (Wang et al., 2021), CTPT (Xu et al., 2023)
Decomposition:
MPT (Wang et al., 2023)

This paper was refined with ChatGPT and Claude.