Make Prompt-based Black-Box Tuning Colorful: Boosting Model Generalization from Three Orthogonal Perspectives

Qiushi Sun[♡], Chengcheng Han[◇], Nuo Chen[◇], Renyu Zhu[☆]
Jingyang Gong[☾], Xiang Li[◇✉], Ming Gao[◇]
[♡]National University of Singapore [◇]East China Normal University,
[☆]NetEase Fuxi AI Lab [☾]New York University
qiushisun@u.nus.edu, {chengchenghan,nuochen}@stu.ecnu.edu.cn
zhurenyu@corp.netease.com, jingyang.gong@nyu.edu
{xiangli,mgao}@dase.ecnu.edu.cn

Abstract

Large language models (LLMs) have shown increasing power on various natural language processing (NLP) tasks. However, tuning these models for downstream tasks usually needs exorbitant costs or is unavailable due to commercial considerations. Recently, black-box tuning has been proposed to address this problem by optimizing task-specific prompts without accessing the gradients and hidden representations. However, most existing works have yet fully exploited the potential of gradient-free optimization under the scenario of few-shot learning. In this paper, we describe BBT-RGB, a suite of straightforward and complementary techniques for enhancing the efficiency and performance of black-box optimization. Specifically, our method includes three plug-and-play components: (1) Two-stage derivative-free optimization strategy that facilitates fast convergence and mitigates overfitting; (2) Automatic verbalizer construction with its novel usage under few-shot settings; (3) Better prompt initialization policy based on instruction search and auto-selected demonstration. Extensive experiments across various tasks on natural language understanding and inference demonstrate the effectiveness of our method. Our codes and data are available at https://github.com/QiushiSun/BBT-RGB.

Keywords: Black-box Language Models, Derivative-free Optimization, Parameter-Efficient Tuning

1. Introduction

Transformer-based Language models (Vaswani et al., 2017) have achieved remarkable improvements among various NLP tasks (Qiu et al., 2020; Lin et al., 2022) in recent years. These models are mainly first pre-trained on a large-scale unsupervised corpus and then fine-tuned on a specific downstream task. However, this paradigm of pre-train and fine-tune face challenges in the era of Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2022; Zhang et al., 2022; Scao et al., 2022; Touvron et al., 2023, *inter alia*). The ever-growing model size leads to a non-stop increase in the cost of tuning, and deploying separate copies of LLMs in real applications becomes exorbitantly expensive. Though recent research on Parameter-Efficient Tuning (Li and Liang, 2021; Lester et al., 2021, *inter alia*) alleviates the problem by tuning a small percentage of parameters while keeping the backbone frozen, the second problem arises: *most LLMs are released as a service, and users can only access them through black-box APIs.* This implies that the aforementioned tuning strategies become less viable owing to the inaccessibility of parameters and gradients, thereby causing a dilemma for downstream applications. Sun et al. (2022c) describe this scenario as Language Model-as-a-Service (LMaaS): Users are unable to tune the model parameters but can accomplish the tasks of interest by finding appropriate prompts with limited examples. Then, Black-Box Tuning (BBT) is proposed as a framework for derivative-free optimization under few-shot settings. Subsequently, BBTv2 (Sun et al., 2022a) was introduced as an improved version that prepends prompts to the hidden states of models, rather than merely injecting prompt tokens at the input layer. Recent works (Chai et al., 2022; Diao et al., 2023; Hou et al., 2023) have also revealed that black-box tuning based on this paradigm is promising.

However, the potential of black-box optimization is still not fully exploited. Previous tuning methods are prone to overfit / fall into local optimum under the scenario of few-shot learning. This phenomenon is triggered by both the characteristics of the Derivative-Free Optimization (DFO) algorithm and the unavailability of pre-trained prompts under few-shot settings. Moreover, they do not fully utilize the information returned by the black-box.

In this paper, we present BBT-RGB, a suite of practical, complementary, and pluggable techniques that further explore the possibility of black-box tuning. We take one step forward in black-box tuning from the following three aspects 1) Employ-

✉ Corresponding author.

LREC-COLING 2024, pages 10958–10969

ing a two-stage DFO strategy for the attenuation of overfitting. 2) Utilizing multiple auto-selected verbalizers to exploit the context further. 3) Combining manual prompt with new search approach for task instructions improvement. Extensive experiments across various NLP downstream tasks demonstrate the superiority of our method. Besides, BBT-RGB can significantly outperform current gradient-based Parameter-Efficient tuning methods (Houlsby et al., 2019; Ben Zaken et al., 2022; Hu et al., 2022; Liu et al., 2022) under few-shot learning. Our main contributions can be summarized as follows:

- We propose a two-stage derivative-free optimization strategy that enables stable convergence of training tunable prompts while effectively mitigating the issue of overfitting.
- To further exploit the LLM's output, we propose a verbalizer selection process to derive multiple appropriate candidates. Moreover, instruction with judiciously selected demonstration is adopted for prompt initialization.
- A wide range of NLP tasks is covered to verify the effectiveness of our approach. By employing our method, optimization[1] under the derivative-free framework can reach comparative performance to full fine-tuning.

2. Preliminaries

In this section, we briefly introduce the basics of LLMs, prompt-based learning, and DFO.

Large Language Models and APIs Large language models (LLMs) (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020) have revolutionized the NLP landscape in the past few years. Given some examples of tasks as input, LLMs can be "prompted" to conduct a wide range of NLP tasks. These huge models are usually released as a service (Brown et al., 2020; Chen et al., 2021; Ouyang et al., 2022), which allows users to interact with the models deployed on the cloud servers through APIs. Unlike some popular open-source LMs (Devlin et al., 2019; Liu et al., 2019) that can be directly utilized by researchers, access to the parameters and gradients of LLMs is restricted due to commercial, ethical, and security concerns.

Prompt-based Learning Prompt-based learning (Liu et al., 2023) transforms an NLP downstream task into a masked language modeling (MLM) task and narrows the discrepancy between pre-training and fine-tuning. Based on the prompt format, prompt-based learning can be categorized into discrete prompts and continuous prompts. Discrete prompts can be designed manually (Brown et al., 2020; Schick et al., 2020) or generated automatically (Gao et al., 2021). Continuous prompts are designed as a sequence of vectors (Qin and Eisner, 2021; Lester et al., 2021) that are usually prepended to the input and optimized by gradients. Recently, Sun et al. (2022c) propose BBT for optimizing prompts under gradient-free settings, as is shown in section 3. We mainly focus on the optimization of continuous prompts under the black-box settings in this paper.

Derivative-free Optimization Derivative-free optimization (DFO) algorithms are capable of solving complex problems without the back-propagation process. DFO generally employs a sampling-and-updating framework (Rios and Sahinidis, 2013; Wierstra et al., 2014; Qian et al., 2016) to improve the solution iteratively. For instance, Covariance Matrix Adaptation Evolution Strategy (Hansen and Ostermeier, 2001; Hansen et al., 2003), namely CMA-ES, is a widely adopted evolutionary algorithm for non-linear non-convex continuous optimization. At each iteration, the algorithm samples new potential solutions from a parameterized distribution model (*e.g.*, multivariate normal distribution). Besides, we have COBYLA algorithm (Constrained Optimization BY Linear Approximation) (Powell, 1994, 1998) that builds a linear approximation model of the objective function and constraints within a trust region, iteratively updating the model based on the progress made in minimizing the objective function.

3. Black-Box Tuning

Given a batch of samples (X, Y) converted with prompt templates and label words, the original derivative-free prompt learning, as introduced by Sun et al. (2022a) first use a set of prompt embeddings p to concatenate the input tokens, creating the prompted input for LLMs with frozen backbones. The prompt $p = p_0 + p_\theta$ consists of the initial prompt $p_0 \in \mathbb{R}^D$, which is manually/randomly selected and a tunable prompt $p_\theta \in \mathbb{R}^D$ that is progressively optimized through a DFO algorithm like CMA-ES (Hansen et al., 2003). DFOs suffer slow convergence on high-dimensional problems, but fortunately, Aghajanyan et al. (2021) discover that PLMs exhibit low-dimensional reparameterization that is as effective for fine-tuning as the full parameter space. This finding indicates that the search space of p_θ can be condensed into an intrinsic dimensionality $z \in \mathbb{R}^d$ $(d \ll D)$ by using a (frozen) random projection matrix $\Pi \in \mathbb{R}^{D\times d}$, such that $p_\theta = \Pi \cdot z$ will significantly decrease the cost

[1]We follow bbtv2 (Sun et al., 2022a) to use random projection matrices to transform prompt parameters into low-dimensional subspaces.

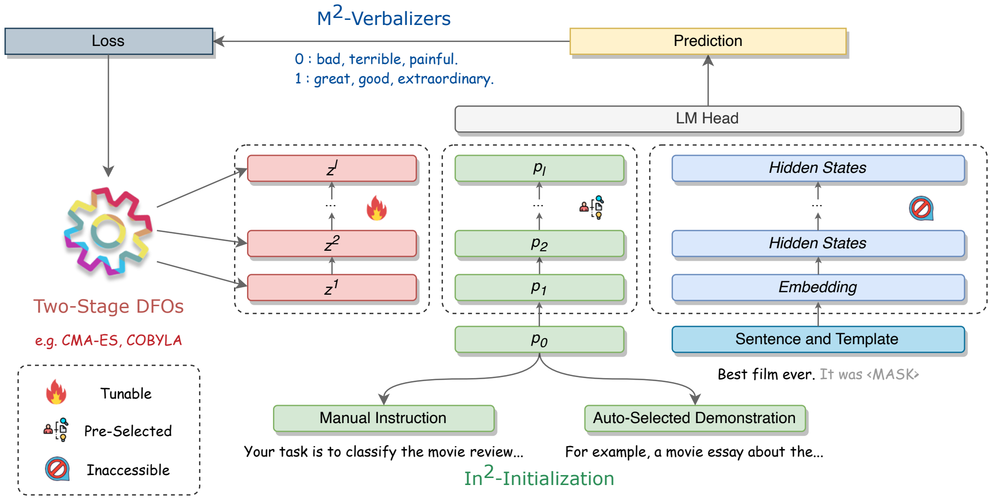


Figure 1: An illustration of BBT-RGB. Given a backbone model with L layers. The target is to optimize continuous prompts $z^l, l \in [1, L]$. We use **R**ed, **G**reen and **B**lue to indicate three distinct aspects of our strategy, which inspired the naming of our method. M[2] Verbalizers (Multi-Mixed Verbalizers) further utilize the information provided by the LLMs. In[2] Initialization (Instruction learning + In-context learning) improves prompt tuning by integrating both instruction and demonstration, noted as p_l. And Two-Stage DFOs exploit the advantages of different optimization methods. ⚙ represents the combination of DFOs.

of optimization. Subsequently, the task-specific inference of model f through API Call is performed to determine the fitness of candidate prompts using an objective function $\mathcal{L}(f([P;X]),Y)$, where $\mathcal{L}$ is a loss function such as cross-entropy. Finally, the DFO algorithm iteratively refines the prompt for seeking $p^* = \arg\min_p \mathcal{L}(f([P;X]),Y)$.

In the era of LLMs, black-box optimization is a promising research target that can drive models for few-shot learning without access to gradients. Sun et al. (2022c) first propose BBT that focuses on optimizing continuous prompt by only accessing inference APIs and then present BBTv2 (Sun et al., 2022a) as an improved version. While some recent works focus on optimizing discrete prompts concurrent with our work. Diao et al. (2023) present black-box discrete prompt learning with gradient estimation as their key feature. Hou et al. (2023) first use gradient-free methods to sample sub-optimal discrete prompts and then ensemble them by boosting algorithm. And Chai et al. (2022) acquire informative feedback to enhance derivative-free optimization through using frozen subnetworks as critics. Recently, Han et al. (2023) ingeniously leverage knowledge distillation to combine gradient descent and gradient-free optimization.

4. BBT-RGB

As is shown in Figure 1, we introduce our method: BBT-RGB, which contains three orthogonal optimization perspectives of derivative-free learning.

4.1. Two-Stage DFOs

Previous works of black-box tuning mainly use CMA-ES to optimize the intrinsic dimensionality (Aghajanyan et al., 2021) of LLMs. Nonetheless, in the early training stage, the evolutionary algorithm (EA) exhibits a considerably faster convergence rate compared to the search-based algorithm (SA), which potentially causes fast overfitting. Then, the following steps would be futile. Thus, we design a novel two-stage DFO algorithm for black-box tuning, as is shown in algorithm 1.

We leverage the advantages of two different kinds of DFOs respectively. In stage I, we use EA to perform coarse-grained population-level optimization, which has a specific budget (Number of API Calls) to move toward the target swiftly. And the SA will use the remaining budgets in stage II for approximating the solution by dimension-level fine-grained search.

4.2. M[2] Verbalizers

Verbalizers, defined as words that can serve as labels, play a crucial role in prompt learning (Schick et al., 2020; Schick and Schütze, 2021a). However, most prior works employ a single verbalizer for gradient-free optimization, which cannot fully use the information, *i.e.*, logits returned by the black box model. To address this problem, we propose **M**ulti-**M**ixed verbalizers, which are constructed through the following methods: 1) manual verbalizer selection[2]. 2) search-based verbalizer construction

[2]Specifically, we use synonyms in practice.

Algorithm 1 Two-Stage DFOs

Input: popsize: λ, intrinsic dimension: d
Input: budget1:$b1$, budget2:$b2$, backbone:f_{model}
Input: m, δ, C, D // initialize state variables
Output: hidden variable: z

```
function TWO-STAGE DFO
  // CMA-ES
  repeat
    for each hidden layer do
      for i in 1 to λ do
        Sample z_i from N(m, δ^2 C)
        f_i = f_model(z_i)
      end for
      Update m, δ, C with f
    end for
    Update z to min(f)
  until b1 times f_model call
  // COBYLA
  for each hidden layer do
    repeat
      for each search direction i in D do
        Update z to min(f) along i
      end for
      Select a new set of D
    until b2//d times f_model call
  end for
end function
```

based on importance estimation by TF-IDF. 3) auto verbalizer generation based on neural nets (Gao et al., 2021). After the aforementioned approaches select verbalizers, the confidence of each category is represented by the average prediction probability of multiple verbalizers. Compared with the previous approach, M^2 verbalizers take one step forward to exploit the information provided by the black box. Additionally, this approach can prevent the negative impact on model performance caused by a single unsuitable label word.

4.3. In^2 Initialization

An appropriate initialization has proven to play an essential role in effective prompt-based tuning while existing BBT methods mainly employ arbitrary selections for instructions and demonstrations (*e.g.*, randomly selecting some tokens from the vocabulary). Inspired by previous efforts (An et al., 2022; Prasad et al., 2022), we propose a model-agnostic strategy named as In^2 initialization. The first component of our approach is a task-specific manual **In**struction. For the second part, we iterate through the training set and take each sample as a demonstration (Min et al., 2022), which is assessed on the validation set together with the pre-selected instruction. After that, the sample with the best performance is selected for **In**-context learning.

5. Experiments

5.1. Experimental Settings

Backbones We use RoBERTa$_{\text{LARGE}}$ (Liu et al., 2019) as backbone throughout the main experiments. To verify the versatility, we also evaluate on other models including GPT-2$_{\text{LARGE}}$, T5$_{\text{LARGE}}$ (Raffel et al., 2020) and BART$_{\text{LARGE}}$ (Lewis et al., 2020).

Datasets. To evaluate our proposed methods, we choose a series of tasks from GLUE (Wang et al., 2018). Specifically, we employ SST-2 (Socher et al., 2013) and Yelp (Zhang et al., 2015) for sentiment analysis, AGNews and DBPedia (Zhang et al., 2015) for topic classification, SNLI (Bowman et al., 2015) and RTE (Dagan et al., 2005) for natural language inference, and MRPC (Dolan and Brockett, 2005) for semantic paraphrasing.

Methods and Hyperparameters. For all the experiments, we adhered to the same settings as Sun et al. (2022a). For the optimization, the API call limit for each DFO algorithm is set to 8000. Regarding the baselines, we employ the results reported by Sun et al. (2022a) for comparison.

5.2. Main Results

As is demonstrated in table 1, we compare BBT-RGB with both gradient-based and gradient-free tuning methods. We observed different levels of improvement on various NLP tasks.

Sentiment Analysis. On both the SST-2 and Yelp datasets, our method surpasses all prior white-box methods, consistently demonstrating superior performance compared to the established baselines.

Topic Classification. Compared with the previous gradient-free method, BBT-RGB has a significant advancement in the evaluation based on DBPedia and AGNews but still needs to catch up to full model tuning. We hold the view that this is caused by a relatively large number of classes (categories), and it is difficult for the model to learn enough knowledge under few-shot settings.

Entailment and Inference. BBT-RGB benefits entailment and natural language inference tasks significantly; both experiments on SNLI and MRPC indicate surpassing full fine-tuning performance. In addition, we can observe a leap in the accuracy of RTE compared with previous baselines.

5.3. Ablation Studies and Analysis

Ablation Studies. We conduct ablation studies to verify the effectiveness of the three proposed techniques that formed the core of this paper, as demonstrated in Table 2.

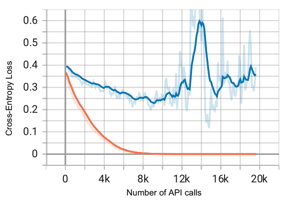
Method	Tunable Params	SST-2 acc	Yelp P. acc	AG's News acc	DBPedia acc	MRPC F1	SNLI acc	RTE acc	Avg.
Gradient-Based Methods									
Model Tuning	355M	85.39 ±2.84	91.82 ±0.79	86.36 ±1.85	97.98 ±0.14	**77.35** ±5.70	54.64 ±5.29	**58.60** ±6.21	**78.88**
Prompt Tuning	50K	68.23 ±3.78	61.02 ±6.65	84.81 ±0.66	87.75 ±1.48	51.61 ±8.67	36.13 ±1.51	54.69 ±3.79	63.46
P-Tuning v2	1.2M	64.33 ±3.05	**92.63** ±1.39	83.46 ±1.01	97.05 ±0.41	68.14 ±3.89	36.89 ±0.79	50.78 ±2.28	70.47
Adapter	2.4M	83.91 ±2.90	90.99 ±2.86	86.01 ±2.18	**97.99** ±0.07	69.20 ±3.58	57.46 ±6.63	48.62 ±4.74	76.31
LoRA	786K	**88.49** ±2.90	90.21 ±4.00	**87.09** ±0.85	97.86 ±0.17	72.14 ±2.23	**61.03** ±8.55	49.22 ±5.12	78.01
BitFit	172K	81.19 ±6.08	88.63 ±6.69	86.83 ±0.62	94.42 ±0.94	66.26 ±6.81	53.42 ±10.63	52.59 ±5.31	74.76
Gradient-Free Methods									
Manual Prompt	0	79.82	89.65	76.96	41.33	67.40	31.11	51.62	62.56
In-Context Learning	0	79.79 ±3.06	85.38 ±3.92	62.21 ±13.46	34.83 ±7.59	45.81 ±6.67	47.11 ±0.63	60.36 ±1.56	59.36
BBT	500	89.56 ±0.25	91.50 ±0.16	81.51 ±0.79	79.99 ±2.95	61.56 ±4.34	46.58 ±1.33	52.59 ±2.21	71.90
BBTv2	12K	90.33 ±1.73	92.86 ±0.62	85.28 ±0.49	93.64 ±0.68	77.01 ±4.73	57.27 ±2.27	56.68 ±3.32	79.01
BBT-RGB (ours)	12K	**92.89** ±0.26	**94.20** ±0.48	**85.60** ±0.41	**94.41** ±0.73	**79.49** ±1.84	**60.71** ±0.66	**61.82** ±1.20	**81.30**

Table 1: Overall comparison between BBT-RGB and other methods (both gradient-based and gradient-free). The results are obtained with the RoBERTa$_{\text{LARGE}}$ backbone in 16-shot (per class) setting.

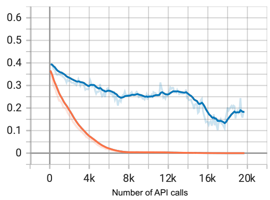
Method	SST-2	AG's News	RTE
BBT-RGB	91.00	85.65	61.80
w/o. M^{2}Verb	91.00	85.59	61.33
w/o. Two-Stage	90.39	85.57	61.17
w/o. In^{2}Init	90.77	84.31	60.17
w/o. Two-Stage & M^{2}Verb	90.83	85.56	59.77
w/o. Two-Stage & In^{2}Init	90.28	83.79	59.30
w/o. In^{2}Init & M^{2}Verb	90.66	83.75	59.47

Table 2: Ablation studies of BBT-RGB on SST-2, AG's News, and RTE

Analysis. We select two cases[3] to analyze the effectiveness of two-stage DFOs on Yelp. In Figure 2, the training loss (orange curve) converges to zero for both methods. While the oscillation of validation loss observed in pure CMA-ES case is mainly attributed to the nature of the adaptive algorithm.



(a) CMA-ES



(b) Two-Stage DFOs

Figure 2: Comparison of original CMA-ES and Two-stage DFOs on Yelp dataset.

In stage II of our proposed two-stage DFOs method, a relatively gentle decrease in validation loss can be observed, demonstrating that dimension-level updates by COBYLA make the overall learning process smoother, which helps us curb the problem of fast overfitting.

[3]We choose CMA-ES (8,000 budgets) and COBYLA (12,000 budgets) for the two-stage DFOs in illustrations.

BBT-RGB Across Models. As shown in Table 3, to demonstrate versatility, we conducted additional experiments on both Decoder-only and Encoder-Decoder architecture models. It is evident that beyond encoder-only models, BBT-RGB also exhibits superior performance on other model architectures.

LM	Method	SST-2	AG's News	DBPedia
Decoder-Only Models				
GPT-2	BBT	75.53 ±1.98	77.63 ±1.89	77.46 ±0.69
	BBTv2	83.72 ±3.05	79.96 ±0.75	91.36 ±0.73
	BBT-RGB	**86.32** ±0.97	**82.01** ±0.81	**93.52** ±1.13
Encoder-Decoder Models				
T5	BBT	89.15 ±2.01	83.98 ±1.87	92.76 ±0.83
	BBTv2	91.40 ±1.17	85.11 ±1.11	93.36 ±0.80
	BBT-RGB	**92.91** ±0.97	**85.50** ±1.32	**93.74** ±0.56
BART	BBT	77.87 ±2.57	77.70 ±2.46	79.64 ±1.55
	BBTv2	89.53 ±2.02	81.30 ±2.58	87.10 ±2.01
	BBT-RGB	**92.63** ±1.43	**82.76** ±1.74	**88.26** ±1.06

Table 3: Comparison of BBT-RGB and baselines on the large versions of GPT-2, BART, and T5.

6. Conclusion

This paper proposes BBT-RGB, a set of practical techniques to drive more powerful derivative-free prompt-based learning. We make improvements from three independent aspects: (1) Two-stage derivative-free optimization algorithms for attenuating overfitting; (2) Versatile verbalizer construction with a robust selection process; (3) Using Instruction learning and demonstrations to exploit in-context information. All the modules are "plug-and-play", and empirical studies across a series of tasks verify the effectiveness of our method.

Limitations and Ethical Consideration

Limitations. Our limitations are threefold:

- Following previous works (Sun et al., 2022c,a), our proposed method lays much emphasis on the optimization of continuous prompts. It can be applied to a majority of open-source Large Language Models (LLMs), but for some commercial models that do not provide loss, logits, or perplexity, the optimization is constrained to remain in the discrete form at the initial layer of the model.

- Since the algorithm is unable to achieve linear convergence, some of the tasks require more API calls, which may lead to extra costs when running on commercial models.
- Given that In^{2}Init and M^{2}Verb involve the search for verbalizers and demonstrations, our method takes a longer execution time compared to BBTv2, requiring approximately 25% additional runtime.

However, the essence of our contributions could be extended to broader scenarios under gradient-free settings, and we leave them as future research.

Ethical Considerations. Our method: BBT-RGB, aims to exploit the potential of black-box tuning further, and the contribution in this paper is fully methodological. Therefore, this contribution has no direct negative social or ethical impacts. Moreover, given that our approach requires significantly less computational resources compared to full-fine tuning, it is poised to contribute positively to the sustainable development of the AI community.

Acknowledgements

This work is supported by Shanghai “Science and Technology Innovation Action Plan” Project (No.23511100700). Our method is also derived from a prize-winning solution of the *First International Algorithm Case Competition: PLM Tuning Track, Guangdong-Hong Kong-Macao Greater Bay Area*. Finally, we thank our anonymous reviewers for their insightful comments and suggestions.

Bibliographical References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7319–7328. Association for Computational Linguistics.

Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. 2022. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Clip-tuning: Towards derivative-free prompt learning with a mixture of rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 108–117, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harri Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways.

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.

Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, LIN Yong, Xiao Zhou, and Tong Zhang. 2023. Black-box prompt learning for pre-trained language models. *Transactions on Machine Learning Research*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Peter I Frazier. 2018. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *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)*, pages 4921–4933, Online. Association for Computational Linguistics.

Chengcheng Han, Liqing Cui, Renyu Zhu, Jianing Wang, Nuo Chen, Qiushi Sun, Xiang Li, and Ming Gao. 2023. When gradient descent meets derivative-free optimization: A match made in black-box scenario. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 868–880, Toronto, Canada. Association for Computational Linguistics.

Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol. Comput.*, 11(1):1–18.

Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195.

Bairu Hou, Joe O'Connor, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. PromptBoosting: Black-box text classification with ten forward passes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13309–13324. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Marie-Anne Lachaux, Baptiste Roziere, Marc Szafraniec, and Guillaume Lample. 2021. Dobf: A deobfuscation pre-training objective for programming languages. In *Advances in Neural Information Processing Systems*, volume 34, pages 14967–14979. Curran Associates, Inc.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *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)*, pages 4582–4597, Online. Association for Computational Linguistics.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv:2103.10385*.

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 pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

M. J. D. Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162.

M. J. D. Powell. 1994. A direct search optimization method that models the objective and constraint functions by linear interpolation. *Advances in Optimization and Numerical Analysis*, pages 51–67.

M. J. D. Powell. 1998. Direct search algorithms for optimization calculations. *Acta Numerica*, 7:287–336.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models.

Hong Qian, Yi-Qi Hu, and Yang Yu. 2016. Derivative-free optimization of high-dimensional non-convex functions by sequential random embeddings. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1946–1952. IJCAI/AAAI Press.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Luis Miguel Rios and Nikolaos V. Sahinidis. 2013. Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2022a. BBTv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianxiang Sun, Xiangyang Liu, Xipeng Qiu, and Xuanjing Huang. 2022b. Paradigm shift in natural language processing. *Machine Intelligence Research*, 19:169–183.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022c. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022, Baltimore, Maryland, USA*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. 2014. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Language Resource References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*. Asian Federation of Natural Language Processing.

Richard Socher and Alex Perelygin and Jean Wu and Jason Chuang and Christopher D. Manning and Andrew Y. Ng and Christopher Potts. 2013. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. ACL.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiang Zhang and Junbo Jake Zhao and Yann LeCun. 2015. *Character-level Convolutional Networks for Text Classification*.

A. Experimental Details

A.1. Implementation

Most of our experiments[4] are conducted with a single NVIDIA GTX 3090 GPU.

[4]Due to the memory requirements, experiments on MRPC and DBPedia datasets are conducted with NVIDIA V100 GPUs.

A.2. BBT-RGB Settings

The details of using BBT-RGB across seven NLP tasks are listed in Table 4. For each task, we report the average performance and standard deviation across three random seeds (42, 50, 66).

A.3. Hyperparameters Settings

The experimental settings in our paper are listed in Table 5. Sigma1 and Sigma2 are the hyperparameters for CMA-ES. Alpha refers to a constant scalar for stretching the distribution of random projection matrices, as is shown in equation 1.

$$\sigma_A = \frac{\alpha\hat{\sigma}}{\sqrt{d}\sigma_z}, \tag{1}$$

$\hat{\sigma}$ is the standard deviation of word embeddings from RoBERTa-Large, and σ_z is the standard deviation of the normal distribution maintained by the CMA-ES algorithm. The random projection matrices are frozen during the whole optimization process.

A.4. Templates and Verbalizers

The templates and verbalizers we employed are listed in Table 6 and Table 7 respectively.

B. Clarifications on Two-Stage DFOs

Here we make some clarification about the choice of DFOs mentioned in Section 4.1, mainly about the advantages of using a search-based algorithm along with evolutionary algorithms for optimization:

- Search Precision: Search-based algorithm (SA) excels in fine-tuned searches due to better local adaptation, unlike evolutionary algorithms (EA) which may falter in detailed adjustments near optima.

- Better Convergence: SA demonstrates superior convergence, especially for local optima, by effectively leveraging minor variations in the search space.

- Parameters: SA offers greater flexibility in parameter tuning, enabling more precise adjustments when approaching the optimal solution.

- Robustness: SA are more resistant to evaluation noise, crucial for stability in later optimization stages, compared to the noise sensitivity of evolutionary algorithms.

	SST-2	Yelp	AGNews	DBPedia	SNLI	RTE	MRPC
Two-Stage DFO	✓	✓	✓	✓	✓	✓	✗
M^2 Verbalizers	✗	✓	✗	✓	✗	✓	✓
In^2 Initialization	✓	✓	✓	✓	✓	✓	✗

Table 4: The details of employing BBT-RGB, ✓ refers to use the given technique on this task, and ✗ vice versa

Task	Budget1 (CMA-ES)	Budget2 (COBYLA)	Alpha	Sigma1	Sigma2
SST-2	7,000	6,000	0.5	0.7	0.7
Yelp	8,000	6,000	0.9	0.4	0.2
AGNews	8,000	6,000	0.1	0.6	0.2
DBPedia	8,000	6,000	0.3	0.2	0.2
SNLI	8,000	6,000	0.5	0.45	0.2
RTE	8,000	6,000	0.5	1	0.2
MRPC	8,000	0	0.3	0.3	0.2

Table 5: Hyperparameter Settings for BBT-RGB in different tasks.

Dataset	**Template**
SST-2	$\langle P \rangle$ $\langle S \rangle$. It was `[MASK]`
Yelp P.	$\langle P \rangle$ $\langle S \rangle$. It was `[MASK]`
AGNews	$\langle P \rangle$ `[MASK]` News: $\langle S \rangle$
DBPedia	$\langle P \rangle$ [Category: `[MASK]`] $\langle S \rangle$
MRPC	$\langle P \rangle$ $\langle S_1 \rangle$? `[MASK]`, $\langle S_2 \rangle$
RTE	$\langle P \rangle$ $\langle S_1 \rangle$? `[MASK]`, $\langle S_2 \rangle$
SNLI	$\langle P \rangle$ $\langle S_1 \rangle$? `[MASK]`, $\langle S_2 \rangle$

Table 6: Prompt templates used in this paper. $\langle P \rangle$ is a sequence of continuous prompt tokens. $\langle S \rangle$ is the original input text.

Dataset	M^2 Verbalizers
SST-2	**Positive**: exciting, all, indeed, ⋯ **Negative**: ridiculous, worse, stupid, ⋯
Yelp P.	**Positive**: addictive, sensational, classic, ⋯ **Negative**: boring, worse, ugly, ⋯
AG's News	**World**: South, China, Africa, ⋯ **Sports**: Athletics, SPORTS, Sporting, ⋯ **Business**: Banking, Manufacturing, Trade, ⋯ **Tech**: Digital, Internet, Tech, ⋯
DBPedia	**Company**: Business, Products, ⋯ **Educational/Institution**: Education, Schools, ⋯ **Artist**: Artists, ⋯ **Athlete**: Profile, ⋯ **Office Holder**: Politics, ⋯ **Mean Of Transportation**: Vehicles, ⋯ **Building**: Architecture, ⋯ **Natural Place**: Lakes, ⋯ **Village**: Rural, ⋯ **Animal**: Animals, Birds, ⋯ **Plant**: Plants, plants, Flowers, ⋯ **Album**: Album, Records, ⋯ **Film**: Movies, Films, ⋯ **Written Work**: Books, Fiction, ⋯
MRPC	**Equivalent**: Finally, Notably, Next, ⋯ **Not Equivalent**: Instead, Although, That, ⋯
RTE	**Yes**: Indeed, So, Wordwide, ⋯ **No**: Also, Now, meanwhile, ⋯
SNLI	**Yes**: Whatever, YES, Regardless, ⋯ **Maybe**: Imagine, Usually, Typically, ⋯ **No**: Besides, Unfortunately, Surprisingly, ⋯

Table 7: Examples of the M^2 Verbalizers used in practice.