

# Revisiting Automated Prompting: Are We Actually Doing Better?

Yulin Zhou<sup>1</sup> Yiren Zhao<sup>2</sup> Ilya Shumailov<sup>3</sup> Robert Mullins<sup>1</sup> Yarin Gal<sup>3</sup>

<sup>1</sup>University of Cambridge   <sup>2</sup>Imperial College London   <sup>3</sup>University of Oxford  
yz709@cam.ac.uk   a.zhao@imperial.ac.uk   ilia.shumailov@chch.ox.ac.uk  
robert.mullins@cl.cam.ac.uk   yarin@cs.ox.ac.uk

## Abstract

Current literature demonstrates that Large Language Models (LLMs) are great few-shot learners, and *prompting* significantly increases their performance on a range of downstream tasks in a few-shot learning setting. An attempt to automate human-led prompting followed, with some progress achieved. In particular, subsequent work demonstrates that automation can outperform fine-tuning in certain  $K$ -shot learning scenarios (Shin et al., 2020; Zhang et al., 2021). In this paper, we revisit techniques for automated prompting on six different downstream tasks and a larger range of  $K$ -shot learning settings. We find that *automated prompting does not consistently outperform simple manual prompting*. Our work suggests that, in addition to fine-tuning, *manual prompting should be used as a baseline* in this line of research.

## 1 Introduction

Transformer-based Large Language Models (LLMs) are now considered foundation models for downstream tasks (Bommasani et al., 2021). The *pre-train then fine-tune* approach achieved state-of-the-art performance on a range of Natural Language Processing (NLP) tasks (Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020). Unfortunately, in many NLP applications, the lack of high-quality labelled training data is a barrier to producing a model with good performance in the pre-train and then fine-tune approach. To address this issue, *prompt-based learning* (Petroni et al., 2019; Schick and Schütze, 2020a,b; Liu et al., 2021a) emerged as a new paradigm for tuning a high-quality, pre-trained LLM in a *few-shot learning* scenario, where only a few samples are available for downstream task learning.

In the prompt-based learning paradigm (Figure 1), an input  $X$  is modified using a template function  $p$ , also known as a prompting function and has one or more placeholders called mask tokens

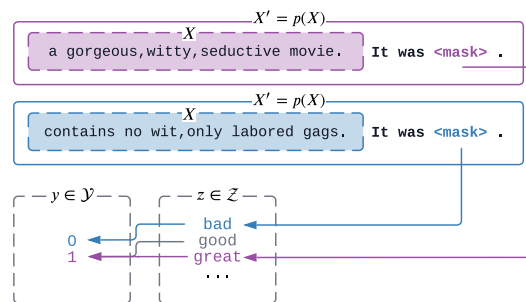


Figure 1: Sentiment analysis with the prompt-based learning paradigm. Input  $X'$  is the prompted input, and there is a many-to-one mapping between answers  $z \in \mathcal{Z}$  and labels  $y \in \mathcal{Y}$ .

$\langle \text{mask} \rangle$ , resulting in a prompted input  $X' = p(X)$  (Liu et al., 2021b). Additionally, a verbaliser designs an answer domain  $\mathcal{Z}$ , so that for an output label domain  $\mathcal{Y}$ , there is a many-to-one mapping for an answer  $z \in \mathcal{V}_y \subseteq \mathcal{Z}$  to an output label  $y \in \mathcal{Y}$  in accordance with the downstream task. Considering a language model  $f_o$  pre-trained on a large corpus of text, such as Wikipedia, the goal of prompt-based learning is to fine-tune it on a small dataset of prompted inputs  $X'$  and corresponding output  $y$ , in order to produce a high-quality language model  $f_p$  capable of generating an answer  $z$  for a given input  $X$ .

Prompting formulates downstream tasks such as sentiment analysis and text classification to cloze completion (also known as filling in the blanks). Furthermore, using prompts and fine-tuning allows models to gain superior few-shot learning capabilities (Lester et al., 2021; Schick and Schütze, 2020a; Shin et al., 2020). Despite the relative success of prompt-based learning, the design of prompts can be a challenging task. As a result, many research studies sought to *automate* the process of designing suitable prompts for downstream tasks (Liu et al., 2021c; Zhang et al., 2021; Shin et al., 2020). The motivation for automating prompt design is usually two-fold: first, manually designing prompts can be

time-consuming; and second, automated ones can often provide better performance. In this work, we question *the second motivation* and demonstrate that *existing automated prompts do not consistently outperform their manual counterparts* under various  $K$ -shot learning setups. In this paper, we make the following contributions:

- We thoroughly investigate automated prompts and demonstrate that they do not consistently outperform manual prompts, even when the latter are created using basic heuristics and selected among a small number of options (Section 3.2).
- We show empirically that fine-tuning only serves a strong baseline when  $K \geq 100$  in a  $K$ -shot learning setup (Section 3.2).
- By visualising the prompts generated by auto-prompting, we explain why these prompts are not necessarily better than manually designed ones (Section 3.4).
- Supported by our empirical evidence and evaluation, we strongly recommend that *future research should consider manual prompts as a simple yet effective baseline*.

## 2 Related Work

The rise of the *prompting-based learning paradigm* comes with the development of LLMs (Brown et al., 2020), which were demonstrated to be good few-shot learners (Liu et al., 2021d). To begin with, researchers focused on manually crafted prompts for downstream tasks (Petroni et al., 2019; Liu et al., 2021b; Scao and Rush, 2021; Zhao et al., 2021; Schick and Schütze, 2020a), yet soon shifted towards automated prompt designs. Schick et al. investigated how to automatically identify label words for a prompt (Schick and Schütze, 2020a,b), while Shin et al. proposed AutoPrompt, a framework for automatically generating prompts for various tasks, through a gradient-based search (Shin et al., 2020). Gao et al. used another LLM, T5 (Raffel et al., 2020), to generate both the prompting templates and verbaliser answer domains (Gao et al., 2020). Han et al. incorporated logic rules into prompt designs, combining several simple sub-prompts according to these rules (Han et al., 2022). All of the above mentioned methods are based on the assumption that the prompt design has to rely on discrete tokens.

Liu et al. and Lester et al. demonstrated that prompts could be trainable continuous embeddings, or soft prompts, instead of discrete tokens. These soft prompts can be learned with a frozen language model (LLM) on a target task (Liu et al., 2021d; Lester et al., 2021; Zhang et al., 2021). Liu et al. further discovered that Deep Prompts, which are soft prompts used in every layer of the model, allow for scaling to large LLMs for complex natural language processing (NLP) tasks (Liu et al., 2021c). Zhang et al. developed Differentiable Prompts, which put the label tokens design of the prompt into a continuous space and optimised it jointly with soft prompts (Zhang et al., 2021). An extensive evaluation was conducted by Zhang et al. on various downstream tasks.

Most of the work on automating prompt design mentioned above has two major motivations: to reduce the amount of time it takes to design prompts manually; and to potentially gain better performance, since manual prompt formats can be sub-optimal (Zhang et al., 2021). While the first motivation may be valid in some cases, it largely depends on the task complexity and the amount of data available – it is sometimes possible for non-experts to design a prompt sufficient for simple tasks with a large amount of data. The principal focus of this work, however, is on the second motivation: *can automated prompts really outperform manual prompts in a consistent manner?* A comparison between automated and manual prompts is lacking in current research. To our knowledge, automated prompting methods focus solely on comparing to fine-tuning in a few-shot learning setup, while a comparisons to manual prompting methods remain unexplored. In this paper, we consider AutoPrompt (Auto) (Shin et al., 2020) and Differential Prompt (Diff) (Zhang et al., 2021) as representatives, where one is based on discrete tokens, while the other is based on continuous embeddings. We compare them with manually designed prompts and fine-tuning without prompting on various tasks.

## 3 Evaluation

### 3.1 Experiment setup

A robust framework was developed to assess prompting model performance under  $K$ -shot learning scenarios where only  $K$  samples per class are available for the training and validation datasets. Three prompting models were re-implemented: LM-BFF (manual) (Gao et al., 2020), AutoPrompt

(Auto) (Shin et al., 2020), and DART (Diff) (Zhang et al., 2021) models. During prompt-based learning, each prompting model is allowed to fine-tune the parameters of the pre-trained language model using the limited training and validation datasets.

### 3.1.1 Datasets and Model

We conducted comprehensive experiments on six datasets to compare the performance of prompting models fine-tuned on the pre-trained RoBERTa-large model (Liu et al., 2019). Table 2 in Appendix B shows we picked three sentiment analysis and three textural entailment tasks.

### 3.1.2 Prompt Templates and Verbalisers

We design prompts to concatenate the input text and the `<mask>` token, alongside a verbaliser that maps from the answer domain to the output label domain. Manually designed prompts and verbalisers are adapted from the Public Pool of Prompts (Bach et al., 2022) and previous work on prompting (Gao et al., 2020; Xu et al., 2022). For each dataset, we selected four to six prompt-and-verbaliser pairs, compared their performance under the same  $K = 16$  few-shot scenario, and picked the best-performing pair for further experiments with different  $K$  values. Detailed manually designed prompts and verbalisers, as well as their performance measures, are illustrated in Table 3, and the best-performing pairs are summarised in Table 4 in Appendix C.

An automated discrete prompt replaces the template with trigger tokens `<T>`. Following the same settings used in AutoPrompt (Shin et al., 2020), we inserted ten trigger tokens between the input text and the `<mask>` token. Under a  $K$ -shot scenario, the verbaliser mapping is automatically generated from the train and validation dataset, each with  $K$  samples per class. Table 5 in Appendix D shows the automated discrete prompts and verbalisers for each dataset. A differential prompt starts from the manually designed prompt but treats both the template and the verbaliser as a collection of differentiable parameters.

Take the dataset SST2 as an example: a suitable manually designed prompt could be “`<sentence>` . It was `<mask>` .” with a verbaliser  $\{\text{bad} \mapsto 0, \text{good} \mapsto 1\}$ ; An automated discrete prompt could be “`<sentence>` `<T>` . . . `<T>` `<mask>` .” with ten trigger tokens `<T>`.

### 3.1.3 Hyper-parameters

We conducted a beam search using the AdamW optimiser (Loshchilov and Hutter, 2017) for the optimal batch size, learning rate and weight decay for each set of experiments with the same dataset and  $K$ -shot value. Each experiment is run with 100 epochs and an early stopping value of 5, *i.e.*, when the validation loss is non-decreasing for 5 epochs. The detailed hyper-parameters used in each set of experiments are listed in Table 6, and details on the evaluation metrics are in Appendix E.

## 3.2 Main Results

Table 1 illustrates the performance of various prompting strategies. We observe that manual prompts exhibit the best performance in 13 out of the 24 setups (6 different datasets and 4 different  $K$ s), and the second-best performance in 8 of them. Automated prompts (both Auto and Diff) only show a clear advantage in TWEETS-HATE-OFFENSIVE when  $K = 100$ . The baseline in Table 1 is direct fine-tuning on the  $K$  samples.

We also see that automated prompts can be catastrophically ineffective in certain setups. For example, as shown in Table 5, Auto performs much worse than Manual or Baseline in MNLI-MATCHED when  $K = 100$ . Diff also significantly underperforms Manual in TWEETS-HATE-OFFENSIVE when  $K = 16$ . In later parts of this section, we provide an analysis of the generated prompts and explore the reasons for this phenomenon. Finally, we demonstrate that Baseline sometimes performs well when  $K$  is large. This is seen in SST2 when  $K = 100, 1000$  and also ENRON-SPAM when  $K = 100$ . In general, we make the following observations:

- Manual prompting outperforms automated prompting (Auto and Diff) with different  $K$ -shot setups on most tasks.
- Automated prompting sometimes cannot even outperform fine-tuning, *e.g.* MNLI-MISMATCHED  $K = 100, 1000$ .
- When  $K$  is small, prompting can greatly improve performance, *e.g.* on SST2 and MNLI.
- Automated prompting can fail catastrophically (*e.g.* MNLI-MISMATCHED  $K = 1000$ ) and have a high variance in performance (*e.g.* 15.5 standard deviation on SST2), while manual prompting is more robust.

$K$	SST2				QNLI			
	Baseline	Auto	Diff	Manual	Baseline	Auto	Diff	Manual
8	59.8 ± 8.6	51.7 ± 1.9	<b>88.0 ± 1.6</b>	77.6 ± 4.6	49.9 ± 1.0	<u>51.5 ± 0.7</u>	50.5 ± 2.1	<b>54.6 ± 2.8</b>
16	72.1 ± 15.0	70.1 ± 3.9	<b>87.8 ± 0.7</b>	<u>86.9 ± 1.6</u>	49.9 ± 0.2	53.4 ± 1.3	<u>59.5 ± 3.6</u>	<b>74.1 ± 1.2</b>
100	<b>89.6 ± 0.5</b>	83.5 ± 4.3	88.6 ± 0.7	<u>89.4 ± 1.0</u>	78.9 ± 2.3	74.0 ± 4.3	<u>80.2 ± 2.1</u>	<b>82.7 ± 0.7</b>
1000	<b>92.7 ± 0.2</b>	<u>92.5 ± 0.2</u>	90.1 ± 0.7	92.3 ± 0.2	<u>87.2 ± 1.0</u>	83.2 ± 3.8	85.2 ± 1.1	<b>88.0 ± 0.3</b>
$K$	MNLI-Matched				MNLI-Mismatched			
	Baseline	Auto	Diff	Manual	Baseline	Auto	Diff	Manual
8	34.6 ± 2.4	34.2 ± 1.1	<u>51.3 ± 1.1</u>	<b>55.7 ± 3.3</b>	33.8 ± 0.8	33.8 ± 0.5	<u>47.6 ± 3.0</u>	<b>56.0 ± 1.4</b>
16	33.3 ± 0.2	34.9 ± 0.7	<b>61.4 ± 1.5</b>	60.2 ± 3.7	32.8 ± 1.3	35.6 ± 0.8	<u>59.4 ± 1.1</u>	<b>60.2 ± 2.7</b>
100	63.1 ± 1.3	42.3 ± 0.5	<u>72.1 ± 0.8</u>	<b>74.1 ± 1.2</b>	<u>73.6 ± 2.1</u>	39.5 ± 1.0	73.3 ± 1.2	<b>77.0 ± 1.2</b>
1000	<u>82.7 ± 0.5</u>	72.9 ± 2.3	80.0 ± 0.8	<b>83.2 ± 0.3</b>	<u>84.3 ± 0.5</u>	76.6 ± 3.7	82.0 ± 0.4	<b>85.0 ± 0.2</b>
$K$	ENRON-SPAM				TWEETS-HATE-OFFENSIVE			
	Baseline	Auto	Diff	Manual	Baseline	Auto	Diff	Manual
8	49.1 ± 36.6	73.4 ± 6.0	<b>80.7 ± 5.7</b>	67.9 ± 12.2	14.5 ± 9.5	12.1 ± 4.6	<b>32.5 ± 7.1</b>	25.8 ± 16.5
16	84.2 ± 4.0	80.5 ± 2.6	88.0 ± 2.3	<b>89.4 ± 3.0</b>	<u>38.0 ± 4.1</u>	42.5 ± 2.6	37.2 ± 7.7	<b>46.7 ± 2.5</b>
100	<b>97.1 ± 0.4</b>	90.8 ± 0.4	96.3 ± 0.8	96.3 ± 0.5	44.9 ± 0.9	51.4 ± 3.4	<b>59.7 ± 2.8</b>	47.0 ± 0.8
1000	98.0 ± 0.5	97.0 ± 0.7	<b>99.0 ± 0.1</b>	<u>98.7 ± 0.2</u>	66.5 ± 1.5	66.8 ± 1.8	<b>67.7 ± 3.3</b>	<u>67.5 ± 2.1</u>

Table 1: The performance of various prompting methods on RoBERTa-large (Liu et al., 2019) was assessed using numbers reported as percentages, with a mean and standard deviation across five independent runs. The best and second-best performing methods are represented in bold and underlined fonts, respectively. The baseline is fine-tuning only without any prompting, while Auto, Diff, and Manual correspond to AutoPrompt (Shin et al., 2020), Differential Prompt (Zhang et al., 2021), and LM-BFF (Gao et al., 2020), respectively.

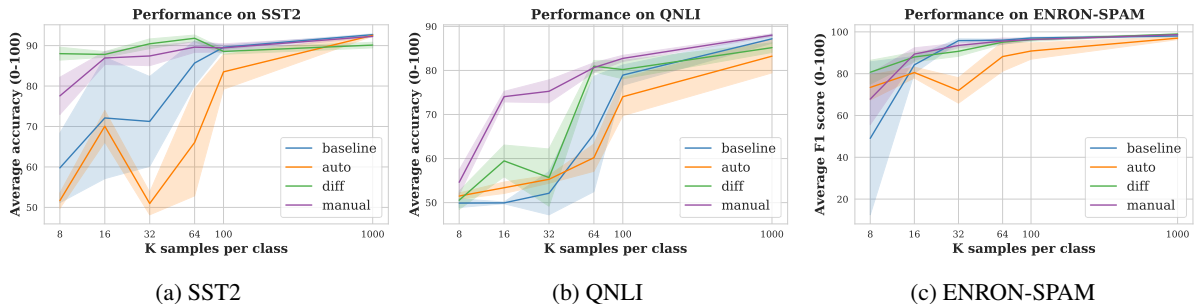


Figure 2: The performance of prompting models on the datasets SST2, QNLI (Wang et al., 2018) and ENRON-SPAM (Metsis et al., 2006) is shown for a wider range of  $K$  values. The solid line plots the mean accuracy across five independent runs, and is bounded by one standard deviation on both sides.

### 3.3 More $K$ -shot Experiments

Figure 2 demonstrates the performance of different prompting styles with more  $K$  values on SST2, QNLI (Wang et al., 2018) and ENRON-SPAM (Metsis et al., 2006).

We observe that the performance of all methods starts to converge with larger  $K$  values, which is consistent with existing literature (Shin et al., 2020). It is also worth mentioning that the automated prompting methods do not consistently outperform manual prompting on this large range of  $K$  values. More results are available in Appendix F.

### 3.4 Visualizing Auto-prompts

As previously discussed, automated prompting can sometimes fail catastrophically. Table 5 sum-

marises all the automated discrete prompts and verbaliser answer domains. Since the answer domain is generated from the  $K$  samples per class, it may not be general enough or optimal for the entire dataset. On the other hand, manual prompts and verbalisers are designed based on common knowledge that humans possess from countless examples encountered in daily life. One possible improvement idea on AutoPrompt is to start with a manually designed prompt and update both the prompt and the verbaliser through a gradient-based search in an iterative manner.

### 3.5 Limitations

All prompting methods are trying to extract knowledge from the Large Language Models (LLMs).



Our paper compares their knowledge extraction abilities. Thus, the performance of RoBERTa-large can serve as a reference point and provide insights for other LLMs. However, it is still necessary to assess each large language model independently to understand its capabilities comprehensively.

We only tested a handful of simple manual prompt-and-verbaliser pairs which are included in Tables 3 and 4. It is entirely possible that there is a lot of room for improvement in the design of manual prompt-and verbaliser pairs, thus providing us a even stronger baseline. We have opted to use ten trigger tokens in Auto, in alignment with the experiment settings originally presented in the AutoPrompt paper (Shin et al., 2020). However, since the verbaliser domains generated under few-shot learning settings are noisy, reducing the number of trigger tokens may improve performance.

## 4 Conclusion

In this paper, we revisit the results generated from automated prompting, and show that *automated prompting cannot consistently outperform simple manual prompting on a variety of tasks*. We also demonstrate that the performance of automated prompting is heavily dependent on the amount of data available, and in some cases can even be worse than fine-tuning. On the other hand, manual prompting is more robust to the amount of data available, and can have similar performance to fine-tuning if not outperforming. We take a closer look at the prompts and verbalisers generated by automated discrete prompting (AutoPrompt) and point out that few-shot learning settings make it challenging to generate prompts and verbalisers that perform well. We hope that this work will motivate researchers to use manual prompts as a general baseline.

## Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful suggestions.

## References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan

Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsources: An integrated development environment and repository for natural language prompts.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv:1703.04009*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021c. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021d. Gpt understands, too. *arXiv preprint 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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes - which naive bayes? In *International Conference on Email and Anti-Spam*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493*.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It’s not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv:1804.07461*.

Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the universal vulnerability of prompt-based learning paradigm.

Ningyu Zhang, Luoqi Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

## A Model and infrastructure details

The RoBERTa-large model (Liu et al., 2019) is pre-trained on a large corpus of raw English text using masked language modelling (MLM) objective; it contains 354 million parameters.

All our experiments are run parallelly on 4 NVIDIA Tesla V100 GPUs; for smaller  $K$  values (e.g.,  $K = 100$ ), most experiments require less than 1 GPU hour, while a setting with a larger  $K$  value (e.g.,  $K = 1000$ ) may require 2 GPU hours.

## B Dataset details

We conducted comprehensive experiments on six datasets (SST2, QNLI, MNLI-MATCHED, MNLI-MISMATCHED, ENRON-SPAM and TWEETS-HATE-OFFENSIVE) to compare the performance of prompting models fine-tuned on the pre-trained RoBERTa-large model. As shown in Table 2, we picked three sentiment analysis and three textural entailment tasks. Among the six, three are binary classifications (SST2, QNLI and ENRON-SPAM), while the remaining datasets have three categories each (MNLI-MATCHED, MNLI-MISMATCHED and TWEETS-HATE-OFFENSIVE).

## C Manual prompt-and-verbaliser designs

In the **Prompt Templates and Verbalisers** part in Section 3.1.2, we discussed how we picked the best-performing prompt-and-verbaliser pairs. We show the picked manual prompt with their picked verbalisers in Table 3, covering SST2, QNLI, MNLI-MATCHED, MNLI-MISMATCHED, ENRON-SPAM and TWEETS-HATE-OFFENSIVE. The underlying mechanism for finding a good manual prompt is detailed in Section 3.1.2. As one can see in these tables, the manual prompts used are very simple and requires minimal domain knowledge.

## D Generated Auto-prompts

In the **Prompt Templates and Verbalisers** part in Section 3.1.2, we also mentioned that an automated discrete prompt replaces the template with trigger tokens  $\langle T \rangle$ . Following the same settings used in AutoPrompt (Shin et al., 2020), we inserted ten trigger tokens between the input text and the  $\langle mask \rangle$  token. All automated discrete prompts and their automatically generated verbalisers are listed in Table 5. In contrast to the manual prompts shown in Appendix C, the auto-prompts generated are now more complex.

## E Hyper-parameters and evaluation metrics for training

In terms of the evaluation metrics which measure the performance of the prompting models,

Dataset	# Class	Test Sample	Description
SST2	2	33674	A sentiment analysis task on movie reviews from the GLUE benchmark (Wang et al., 2018). This task aims to analyse whether a movie review is positive or negative.
QNLI	2	5463	A textual entailment task on question-answer pairs from the GLUE benchmark (Wang et al., 2018). The objective is to determine whether the context sentence contains the answer to the question.
MNLI-MATCHED	3	4907	A multi-class (i.e., entailment, neutral, contradiction) textual entailment task on premise-hypothesis pairs from the GLUE benchmark (Wang et al., 2018). Matched version only preserves pairs within the same genre (e.g., science fiction, speech).
MNLI-MISMATCHED	3	4916	Same as MNLI-MATCHED, the mismatched version is a textual entailment task on premise-hypothesis pairs from the GLUE benchmark (Wang et al., 2018), but it only preserves pairs within different genres.
ENRON-SPAM	2	15858	A safety critical binary sentiment analysis task determining whether an email text is a spam (Metsis et al., 2006).
TWEETS-HATE-OFFENSIVE	3	12391	A safety critical multi-class sentiment analysis task which aims to classify whether a tweet text contains hate speech, offensive speech or neither (Davidson et al., 2017).

Table 2: Six datasets selected in the project. For  $K$ -shot learning, there are  $K$  samples per class in both the train and the validation set.

Prompt Design	SST2		QNLI		
	Answer $\mapsto$ Label	Accuracy	Prompt Design	Answer $\mapsto$ Label	Accuracy
<sentence> . It was <mask> .	terrible $\mapsto$ 0, great $\mapsto$ 1	86.0 $\pm$ 2.7	<question> ? <mask> , <sentence> .	{Yes $\mapsto$ 0, No $\mapsto$ 1}	64.5 $\pm$ 4.8
	bad $\mapsto$ 0, good $\mapsto$ 1	<b>86.9 <math>\pm</math> 1.6</b>	<question> . <mask> , <sentence> .		60.5 $\pm$ 2.3
	dog $\mapsto$ 0, cat $\mapsto$ 1	84.7 $\pm$ 3.4	<question> ? <mask> <sentence> .		68.7 $\pm$ 3.2
	cat $\mapsto$ 0, dog $\mapsto$ 1	68.7 $\pm$ 6.9	<sentence> ? <mask> , <question> .		<b>74.1 <math>\pm</math> 1.2</b>
	great $\mapsto$ 0, terrible $\mapsto$ 1	67.4 $\pm$ 5.0	<question> <mask> <sentence>		50.0 $\pm$ 0.2
			<sentence> ? <mask> , <question>		66.7 $\pm$ 10.2
Prompt Design	MNLI-Matched		MNLI-Mismatched		
	Answer $\mapsto$ Label	Accuracy	Prompt Design	Answer $\mapsto$ Label	Accuracy
<premise> ? <mask> , <hypothesis> . <premise> . <mask> , <hypothesis> . <premise> ? <mask> <hypothesis> . <hypothesis> ? <mask> , <premise> . <premise> <mask> <hypothesis> <hypothesis> ? <mask> , <premise>	{Yes $\mapsto$ 0, Maybe $\mapsto$ 1, No $\mapsto$ 2}	<b>60.2 <math>\pm</math> 3.7</b>	<premise> ? <mask> , <hypothesis> .	{Yes $\mapsto$ 0, Maybe $\mapsto$ 1, No $\mapsto$ 2}	<b>60.2 <math>\pm</math> 2.7</b>
		58.6 $\pm$ 4.8	<premise> . <mask> , <hypothesis> .		56.3 $\pm$ 1.5
		55.6 $\pm$ 1.7	<premise> ? <mask> <hypothesis> .		58.4 $\pm$ 1.1
		51.9 $\pm$ 4.2	<hypothesis> ? <mask> , <premise> .		57.9 $\pm$ 0.8
		51.2 $\pm$ 4.2	<premise> <mask> <hypothesis>		49.4 $\pm$ 2.4
		52.4 $\pm$ 2.9	<hypothesis> ? <mask> , <premise>		56.0 $\pm$ 1.0
Prompt Design	ENRON-SPAM		TWEETS-HATE-OFFENSIVE		
	Answer $\mapsto$ Label	F1 score	Prompt Design	Answer $\mapsto$ Label	F1 score
<mask> : <text> . This is a <mask> : <text> . <mask> email : <text> . <text> . This was a <mask> .	ham $\mapsto$ 0, spam $\mapsto$ 1	82.8 $\pm$ 1.9	<tweet> . This post is <mask> .	{hateful $\mapsto$ 0, offensive $\mapsto$ 1, harmless $\mapsto$ 2}	<b>46.7 <math>\pm</math> 2.5</b>
	ham $\mapsto$ 0, spam $\mapsto$ 1	82.8 $\pm$ 2.8	This post is <mask> : <tweet> .		40.3 $\pm$ 3.8
	genuine $\mapsto$ 0, spam $\mapsto$ 1	<b>89.4 <math>\pm</math> 3.0</b>	<tweet> . This was <mask> .		39.8 $\pm$ 4.5
	ham $\mapsto$ 0, spam $\mapsto$ 1	76.8 $\pm$ 3.3	<mask> speech : <tweet> .		36.8 $\pm$ 11.7

Table 3: The prompt-and-verbaliser pairs are tested under the few-shot scenario  $K = 16$ , and the best-performing pair is highlighted in bold. The mean and standard deviation of scores are computed across five independent runs.

Dataset	Prompt Design	Answer $\mapsto$ Label
SST2	<sentence> . It was <mask> .	bad $\mapsto$ 0, good $\mapsto$ 1
QNLI	<sentence> ? <mask> , <question> .	Yes $\mapsto$ 0, No $\mapsto$ 1
MNLI-MATCHED	<premise> ? <mask> , <hypothesis> .	Yes $\mapsto$ 0, Maybe $\mapsto$ 1, No $\mapsto$ 2
MNLI-MISMATCHED	<premise> ? <mask> , <hypothesis> .	Yes $\mapsto$ 0, Maybe $\mapsto$ 1, No $\mapsto$ 2
ENRON-SPAM	<mask> email : <text> .	genuine $\mapsto$ 0, spam $\mapsto$ 1
TWEETS-HATE-OFFENSIVE	<tweet> . This post is <mask> .	hateful $\mapsto$ 0, offensive $\mapsto$ 1, harmless $\mapsto$ 2

Table 4: Summarised for each dataset, the best-performing manual prompt and verbaliser.

Task	Prompt design	$K$	Answer $\mapsto$ Label
SST2	<sentence> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <mask> .	8	impunity $\mapsto$ 0, ASHINGTON $\mapsto$ 1
		16	worthless $\mapsto$ 0, Kom $\mapsto$ 1
		32	Worse $\mapsto$ 0, $\hat{\text{A}}\text{C}\text{E}$ $\mapsto$ 1
		64	horrible $\mapsto$ 0, magic $\mapsto$ 1
		100	worse $\mapsto$ 0, $\hat{\text{A}}\text{C}\text{E}$ $\mapsto$ 1
		1000	worse $\mapsto$ 0, Excellent $\mapsto$ 1
QNLI	<question> <mask> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <sentence>	8	implement $\mapsto$ 0, defensively $\mapsto$ 1
		16	counter $\mapsto$ 0, Bits $\mapsto$ 1
		32	Meteor $\mapsto$ 0, univers $\mapsto$ 1
		64	ormon $\mapsto$ 0, stood $\mapsto$ 1
		100	idelines $\mapsto$ 0, opard $\mapsto$ 1
		1000	G $\mapsto$ 0, overloaded $\mapsto$ 1
MNLI-MATCHED	<premise> <mask> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <hypothesis>	8	efforts $\mapsto$ 0, democratically $\mapsto$ 1, Congratulations $\mapsto$ 2
		16	OWN $\mapsto$ 0, hypocritical $\mapsto$ 1, examiner $\mapsto$ 2
		32	Alicia $\mapsto$ 0, historians $\mapsto$ 1, BF $\mapsto$ 2
		64	tweets $\mapsto$ 0, onboard $\mapsto$ 1, Anniversary $\mapsto$ 2
		100	filmmakers $\mapsto$ 0, combat $\mapsto$ 1, absence $\mapsto$ 2
		1000	thus $\mapsto$ 0, MED $\mapsto$ 1, independent $\mapsto$ 2
MNLI-MISMATCHED	<premise> <mask> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <hypothesis>	8	Whilst $\mapsto$ 0, oka $\mapsto$ 1, smokers $\mapsto$ 2
		16	Accordingly $\mapsto$ 0, )? $\mapsto$ 1, foreigners $\mapsto$ 2
		32	ibliography $\mapsto$ 0, qa $\mapsto$ 1, Governments $\mapsto$ 2
		64	LER $\mapsto$ 0, jack $\mapsto$ 1, foreigners $\mapsto$ 2
		100	HEL $\mapsto$ 0, gaming $\mapsto$ 1, imperialism $\mapsto$ 2
		1000	Vladimir $\mapsto$ 0, acting $\mapsto$ 1, dislike $\mapsto$ 2
ENRON-SPAM	<question> <mask> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <sentence>	8	Reviewer $\mapsto$ 0, Pure $\mapsto$ 1
		16	debian $\mapsto$ 0, Discount $\mapsto$ 1
		32	hillary $\mapsto$ 0, Vampire $\mapsto$ 1
		64	schedules $\mapsto$ 0, Romance $\mapsto$ 1
		100	subcommittee $\mapsto$ 0, Beauty $\mapsto$ 1
		1000	committee $\mapsto$ 0, ophobic $\mapsto$ 1
TWEETS-HATE-OFFENSIVE	<premise> <mask> <T> <T> <T> <T> <T> <T> <T> <T> <T> <T> <hypothesis>	8	Slater $\mapsto$ 0, herself $\mapsto$ 1, issued $\mapsto$ 2
		16	kicking $\mapsto$ 0, her $\mapsto$ 1, selections $\mapsto$ 2
		32	athi $\mapsto$ 0, herself $\mapsto$ 1, vernight $\mapsto$ 2
		64	racist $\mapsto$ 0, Marie $\mapsto$ 1, skies $\mapsto$ 2
		100	racist $\mapsto$ 0, vaginal $\mapsto$ 1, Miracle $\mapsto$ 2
		1000	homophobia $\mapsto$ 0, b***h $\mapsto$ 1, heavens $\mapsto$ 2

Table 5: Auto prompts designed alongside with the automatically generated verbalisers for each dataset.

Dataset	Model	Batch Size	$\eta$	$w_d$	Dataset	Model	Batch Size	$\eta$	$w_d$
SST2	Auto	8	1e-5	0.01	QNLI	Auto	4	2e-5	0.1
	Diff	8	1e-5	0.01		Diff	4	1e-5	0.1
	Manual	4	2e-5	0.01		Manual	4	2e-5	0.01
MNLI-MATCHED	Auto	4	2e-5	0.01	MNLI-MISMATCHED	Auto	4	2e-5	0.01
	Diff	4	1e-5	0.01		Diff	8	1e-5	0.01
	Manual	4	2e-5	0.01		Manual	4	2e-5	0.01
ENRON-SPAM	Auto	8	1e-5	0.01	TWEETS-HATE-OFFENSIVE	Auto	8	2e-5	0.1
	Diff	8	2e-5	0.0		Diff	8	2e-5	0.0
	Manual	8	2e-5	0.05		Manual	8	2e-5	0.1

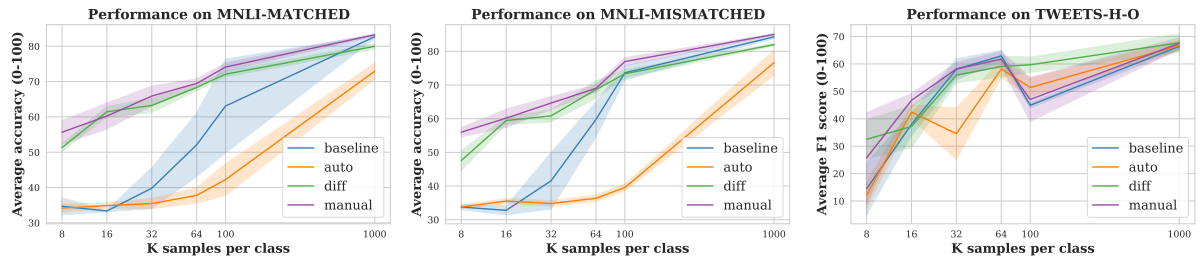
Table 6: Details of the selected hyper-parameters, including batch size, learning rate  $\eta$  and weight decay  $w_d$  for each set of experiments with the same dataset and prompting model.

we utilised two different metrics according to the nature of the datasets: (1) Multi-class classification accuracy for balanced datasets SST2, QNLI, MNLI-MATCHED and MNLI-MISMATCHED. (2) F1 score captures both precisions and recalls for safety-critical or unbalanced datasets ENRON-SPAM and TWEETS-HATE-OFFENSIVE.

Table 6 provides details for the training setups. We show the batch sizes, learning rates and weight decay values used in the experiments. We also

show the optimal hyper-parameters for each set of experiments with the same dataset and prompting model. For example, the optimal hyper-parameters for the dataset SST2 with the prompting model Auto are batch size 8, learning rate  $10^{-5}$  and weight decay 0.01.





(a) MNLI-MATCHED (b) MNLI-MISMATCHED (c) TWEETS-HATE-OFFENSIVE  
 Figure 3: The performance of prompting models on MNLI-MATCHED, MNLI-MISMATCHED (Wang et al., 2018) and TWEETS-HATE-OFFENSIVE (Davidson et al., 2017) is shown for a wider range of  $K$  values. The solid line plots the mean accuracy across five independent runs, and is bounded by one standard deviation on both sides.

## F Additional results for more K-shot experiments

In Figure 2 (Section 3.3), we show the performance with more  $K$  values for SST2, QNLI and ENRON-SPAM. Additional results in the same setup are shown in Figure 3.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Section 3.5*
- A2. Did you discuss any potential risks of your work?  
*Not applicable. Left blank.*
- A3. Do the abstract and introduction summarize the paper’s main claims?  
*Abstract and Section 1*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Section 3*

- B1. Did you cite the creators of artifacts you used?  
*Section 3.1*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Appendix A and Appendix B*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Section 3, Appendix A and Appendix B*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. Left blank.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Section 3, Appendix A and Appendix B*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Appendix B*

### C Did you run computational experiments?

*Section 3*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Appendix A*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Section 3.1 and Appendix E*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Section 3.2 3.3 and Appendix F*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Section 3.1 and Appendix A*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*No response.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*No response.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*No response.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*No response.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*No response.*