Unveiling the Lexical Sensitivity of LLMs: Combinatorial Optimization for Prompt Enhancement

Pengwei Zhan[♢♣♡], **Zhen Xu**[♢], **Qian Tan**[♢*], **Jie Song**[♢♣], **Ru Xie**[♢♣]
[♢]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[♣]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[♡]Sangfor Technologies Inc.
`{zhanpengwei,xuzhen,tanqian,songjie,xieru}@iie.ac.cn`

Abstract

Large language models (LLMs) demonstrate exceptional instruct-following ability to complete various downstream tasks. Although this impressive ability makes LLMs flexible task solvers, their performance in solving tasks also heavily relies on instructions. In this paper, we reveal that LLMs are over-sensitive to lexical variations in task instructions, even when the variations are imperceptible to humans. By providing models with neighborhood instructions, which are closely situated in the latent representation space and differ by only one semantically similar word, the performance on downstream tasks can be vastly different. Following this property, we propose a black-box **C**ombinatorial **O**ptimization framework for **P**rompt **L**exical **E**nhancement (COPLE). COPLE performs iterative lexical optimization according to the feedback from a batch of proxy tasks, using a search strategy related to word influence. Experiments show that even widely-used human-crafted prompts for current benchmarks suffer from the lexical sensitivity of models, and COPLE recovers the declined model ability in both instruct-following and solving downstream tasks.

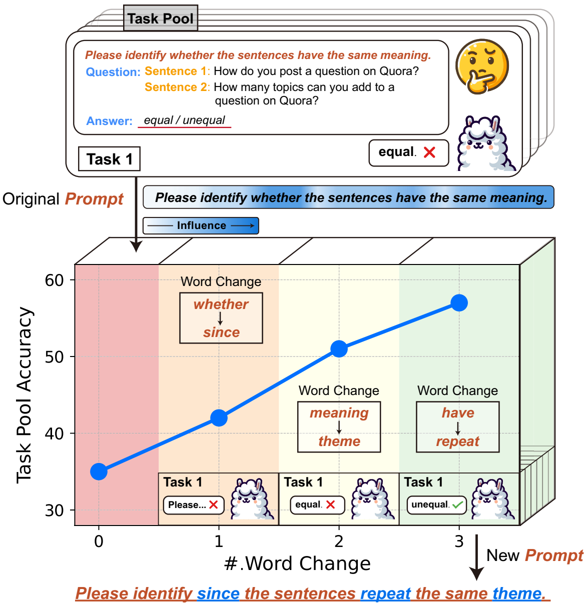


Figure 1: Prompt lexical enhancement from a combinatorial optimization perspective. Initially, we provide the prompt "*Please identify whether the sentences have the same meaning*" for `Llama-2-7B-chat` to complete the tasks from Quora Question Pairs2 (QQP), and combine the validation set of QQP with the prompt as a predefined task pool, with each example being an individual task. By iteratively substituting the most influential words in the prompt with semantically similar words picked from the potential search space, we find the optimal prompt "*Please identify since the sentences repeat the same theme*" that increases the accuracy from 35% to 57%. The details of operations can be found in §3.3.

1 Introduction

Language models have achieved remarkable performance in recent years, particularly those referred to as large language models (LLMs), which contain scaled-up parameters and size (Kaplan et al., 2020; Brown et al., 2020; Hoffmann et al., 2022; OpenAI, 2022; Touvron et al., 2023; Jiang et al., 2023). These models demonstrate an exceptional ability to follow human instructions and complete downstream tasks after instruction tuning (Ouyang et al., 2022). In contrast to masked language models (MLMs) like BERT (Devlin et al., 2019), LLMs do not require the addition and training of extra layers on top of the pre-trained base model to adapt to different downstream tasks. Instead, they complete a wide range of tasks in the same way of generating text, by following different task instructions.

Although the instruction-following ability of LLMs makes them flexible task solvers, their performance on solving tasks also significantly depends on the instructions (i.e., prompts), which are mainly designed by human intuitively and empirically (Wei et al., 2022; Lu et al., 2022; Kojima et al., 2022; Zhou et al., 2023a). These manually

[*] Corresponding Author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5128–5154
November 12-16, 2024 ©2024 Association for Computational Linguistics

designed prompts that incorporated with human knowledge effectively improve the model's performance on specific tasks. However, following Gao et al. (2018), Garg and Ramakrishnan (2020) and Feng et al. (2018), even a minor lexical modification in the input that is imperceptible to humans can lead to vastly different model attention and outputs. Therefore, it is natural to wonder: *whether the prompts carefully constructed by humans maximize LLMs' performance on downstream tasks?* For example, in the context of a sentiment classification task, while humans may confidently assert that the prompt "*Please classify the sentiment of the given text*" outperforms "*Check the given text*", it is hard to say whether it would outperform a prompt like "*Please analyze the sentiment of the given text*".

The unexpected sensitivity of language models to these imperceptible lexical perturbations suggests the possible existence of an alternate prompt, which differs from the original prompt by only a few substituted words, yet yields superior performance on downstream tasks. This insight allows us to frame the process of discovering such an optimal prompt as a combinatorial optimization problem (Blair, 1990), which consists of two key components: the *search space* and the *search method*. The search space can be defined as the set of all potential substitutions for each word in the original prompt, while the search method specifies the strategy for exploring this space and identifying the optimal substitutions. Figure 1 provides a more intuitive example of the process of finding the optimal prompt from a lexical combinatorial optimization perspective. We argue that even without the complex prompt engineering, minor lexical modifications to prompts yield substantial improvements to a model's performance.

In this paper, we reveal the notable sensitivity of LLMs to lexical variations in prompts, which potentially undermine the effectiveness of human-crafted prompts, from the view of combinatorial optimization. Based on our findings, we also propose a black-box **C**ombinatorial **O**ptimization framework for **P**rompt **L**exical **E**nhancement (COPLE). We summarize our main contributions as follows:

1. We intuitively reveal the notable sensitivity of LLMs to the lexical choices in prompts, which suggests the existence of prompts that, while highly similar to the original, can lead to improved performance on downstream tasks.

2. We propose COPLE, a black-box combinatorial optimization framework that enhances prompts on downstream tasks, which performs iteratively lexical optimization under the guidance of word influence.

3. We evaluate COPLE on popular datasets, models, and initial prompts. The results show that COPLE effectively maximizes the performance of prompts in downstream tasks with only lexical modifications, without accessing model parameters or involving complex prompt engineering with human participation.

2 Related Work

Sensitivity to Imperceptible Changes. The outstanding performance of language models seems to be built upon their excellent understanding of text (Devlin et al., 2019; Dong et al., 2019; Radford et al., 2018a,b; Brown et al., 2020). However, previous works reveal that even imperceptible input perturbations, which do not affect human comprehension, can lead to significant changes in the model's output (Goodfellow et al., 2015; Papernot et al., 2016; Zhan et al., 2023a,b; Carlini and Wagner, 2017). This property has been widely exploited to create adversarial examples, where small modifications to the embedding or input text can cause the model to generate incorrect answers (Gao et al., 2018; Zhan et al., 2022a,b, 2024; Li et al., 2019, 2020; Zang et al., 2020). Therefore, we believe even humans experienced in designing prompts may overlook the performance discrepancies caused by such imperceptible changes.

Prompt Tuning and Optimizing. Similarly, recent efforts to optimize prompts for LLMs find that not only the content (Kojima et al., 2022; Yang et al., 2023) but also the format (Zhou et al., 2023a; Lu et al., 2022; Wei et al., 2023; Madaan and Yazdanbakhsh, 2022; Prasad et al., 2023) of the prompt, such as the order of examples and phrases, significantly influence the model performance. Consequently, this sensitivity of language models to minor changes makes the optimal prompt found by the community increasingly complex. For example, Xu et al. (2023) transforms a prompt of length 6 into one exceeding 900 tokens. However, we argue that also due to this sensitivity, complex variations of the prompt should not be the first operation in prompt optimization, as the performance could be inadvertently constrained by the specific words employed in the prompt. This proposition distinguishes our work from previous studies on prompt optimization: given a prompt that has proven initially effective, we focus on the lexical influence of the prompt on model performance, attempting to *recover* the potential performance drop caused by lexical choices, rather than *creating* a prompt that yields optimal results from scratch (Shin et al., 2020; Zhou et al., 2023b; Zhang et al., 2022; Yang et al., 2023; Prasad et al., 2023).

3 Methodology

3.1 Prompt Enhancement

Suppose we are given a data distribution $\mathcal{D}$ over a sequence of downstream tasks $\mathcal{Z} = \{\mathcal{X}, \mathcal{Y}\}$, and each task in the entire task set can be seen as a pair of question and answer $\{\boldsymbol{X}, \boldsymbol{Y}\}$ that both consist of multiple tokens.To recognize a pre-trained autoregressive language model $f_{\boldsymbol{\theta}}$ as the task solver on the task set $\mathcal{Z}$, we hope it can map the input questions to the output answers $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$. However, as model $f_{\boldsymbol{\theta}}$ is not fine-tuned for a specific task, this mapping can only be held with the help of a task-specific prompt $\boldsymbol{P}_{\mathcal{Z}}(\boldsymbol{X}) = (\boldsymbol{D}, \boldsymbol{E}', \boldsymbol{X}, \boldsymbol{V})$, where $\boldsymbol{D}$ is the task description, $\boldsymbol{E}'$ are optional demo examples for few-shot learning, and $\boldsymbol{V}$ is the verbalizer that limits the responses of the model to a set of label words. We can then formulate the performance of the model on the task set as:

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}[\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{P}_{\mathcal{Z}}(\boldsymbol{X})), \boldsymbol{Y})] \tag{1}$$

where $\mathcal{L}$ is a task-specific loss function that measures the discrepancy between the model's output and the ground truth answer. Following this, we can find that directly optimizing the prompt $\boldsymbol{P}_{\mathcal{Z}}(\boldsymbol{X})$ in the discrete token space is challenging due to the non-differentiable nature of text and the large search space. Therefore, it is more suitable to frame the process of optimizing the prompt as a combinatorial optimization problem, where we aim to find the optimal combination of tokens from a predefined search space that consists of candidate tokens.

In this paper, to be more specific, we focus on investigating the influence of minor lexical changes on the *task description* part of the prompt. Let $\boldsymbol{D} = (d_1, d_2, \ldots, d_n)$ be the sequence of tokens in the task description, and $\mathcal{C}_i \in \mathcal{C}$ denote the search space of token d_i. The optimal alternative task description that recovers the potential performance drop caused by wording can thus be formulated as:

$$\begin{aligned} &\boldsymbol{D}^* = (d_1^*, d_2^*, \ldots, d_n^*), \\ &\quad \text{s.t.} \quad d_i^* \in \mathcal{C}_i \\ &\quad \text{and} \quad \forall i \in \{1, \ldots, n\}, \ \ \Delta d_i < \delta \end{aligned} \tag{2}$$

where Δd_i denotes the difference between d_i and d_i^*, and δ denotes a small maximum allowed difference between them, which limits the possible candidates in the search space to tokens that are semantically similar to the original one. This optimal task description $\boldsymbol{D}^*$ is expected to minimize the expected loss on downstream tasks:

$$\boldsymbol{D}^* = \arg\min_{\boldsymbol{D}} \mathop{\mathbb{E}}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}} [\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{P}_{\mathcal{Z}}(\boldsymbol{X})), \boldsymbol{Y})] \tag{3}$$

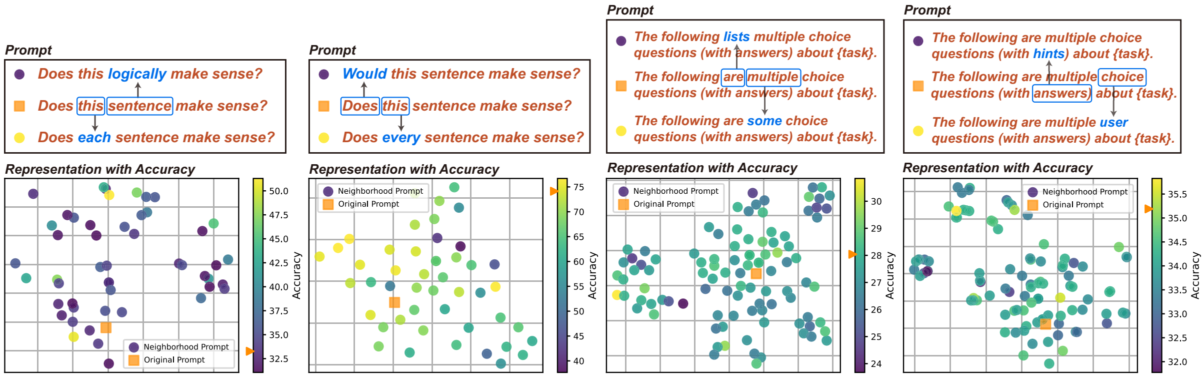
Therefore, the optimal prompt for task $\mathcal{Z}$ can be formulated as $\boldsymbol{P}_{\mathcal{Z}}^*(\boldsymbol{X}) = (\boldsymbol{D}^*, \boldsymbol{E}', \boldsymbol{X}, \boldsymbol{V})$.

3.2 Impact of Minor Lexical Changes

The analysis presented in the previous section relies on a crucial premise: *imperceptible lexical changes in prompts can significantly affect the model performance on downstream tasks*. Before further explaining our approach, we first try to show the validity of this assumption.

Given an initially proven effective prompt, a model, and a predefined task pool, we attempt to demonstrate how prompts within the neighborhood of the original prompt influence the model's performance on the task pool. In this context, we broadly define a prompt's neighborhood as prompts that differ from the original by only *one word* while maintaining a similar meaning. For instance, we consider "*Does the sentence make sense?*" to be within the neighborhood of "*Does this sentence make sense?*". To obtain these qualifying prompts, we employ a MLM. First, we iteratively replace each word in the task description with a `[MASK]` token. We then expect the MLM, by understanding the context, to provide the most probable fill-in words at each position based on the entire task description. After replacing the original words with the fill-in words at each position, we obtain a series of prompts within the neighborhood of the original prompt. We then evaluate the performance of each resulting prompt on the task pool.

Following these definitions and operations, we employ the validation sets of CoLA (Warstadt et al., 2019) and MMLU-STEM (Hendrycks et al., 2021) subtasks, respectively, as predefined task pools. We use `Llama-2-7B-chat` (Touvron et al., 2023) and `Mistral-7B-Instruct-v0.1` (Jiang et al., 2023) as target models and use RoBERTa (Liu et al., 2019) for obtaining neighborhood prompts. The initial prompts are picked from *lm-evaluation-harness* (Gao et al., 2023), and we generate ten most probable fill-in words for each position in



(a) Llama-2 - CoLA (b) Mistral - CoLA (c) Llama-2 - MMLU-STEM (d) Mistral - MMLU-STEM

Figure 2: The visualization of model performance on CoLA and MMLU-STEM validation set with neighborhood prompts. The task description of the original prompt picked for CoLA is "*Does this sentence make sense?*", and for MMLU-STEM is "*The following are multiple choice questions (with answers) about {task}*", where *{task}* is a placeholder to replace with detailed subset type, e.g., "abstract algebra". The point ● in lighter color indicates better performance, and the square 🟧 indicates the original prompt, with the ▶ in the color bar indicating the original performance. The words in the upper prompts indicate the changed words, and words indicate the substitutions.

task description. We then obtain their sentence representations in the target model and project them into a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008). Figure 2 shows the performance of neighborhood prompts on downstream tasks with the distribution of their sentence representations. We can then reach several conclusions on the impact of minor lexical changes in prompts on downstream performance.

1.Semantically similar prompts have vastly different performances on downstream tasks, even if they differ by only one word. For example, when using neighborhood prompts on MMLU-STEM and `Llama-2-7B-chat` (Figure 2(c)), their performance differences can reach 7%. Specifically, changing "*The following are multiple choice questions (with answers) about {task}.*" (🟧) to "*The following lists multiple choice questions (with answers) about {task}.*" (●) reduces the accuracy from the original 28.04% to 23.92%, while changing it to "*The following are some choice questions (with answers) about {task}.*" (●) increases the accuracy to 30.86%. Intuitively, such minor lexical variations should have a minimal impact on semantics, and human performance would likely remain consistent when completing downstream tasks guided by these three prompts (Adam Drewnowski, 1978; McCusker et al., 1981; van Orden, 1987; Rayner et al., 2006). However, models exhibit a high degree of sensitivity to these changes.

2.In the latent representation space, prompts that are in close proximity may have vastly different performance on downstream tasks. In most cases, the performance of neighborhood prompts on downstream tasks does not demonstrate a clear correlation with the distribution of their sentence representations. Even when the representations of prompts are clustered together, they can still have substantial performance discrepancies. For example, in the representation space of `Llama-2-7B-chat`, the best-performing prompt (51.2%, ●) on CoLA (Figure 2(a)) is situated in very close proximity to the prompt with nearly the worst performance (32.4%, ●). From the perspective of sentence representations, this observation indicates that even for semantically highly similar prompts, their performance may be vastly different, and it is difficult to infer their performance from one another directly.

3.3 COPLE

According to our findings, even for semantically similar prompts with only one word difference, their performance on downstream tasks may be very different, and we cannot infer the performance of one prompt from another seemingly similar prompt. Therefore, we propose COPLE, trying to recover the degraded ability of models caused by lexical sensitivity. The key idea behind COPLE is to guide the lexical optimization of the initial prompt by the model performance on a batch of reference tasks i.i.d. to the downstream tasks, and iteratively improve the prompt based on the feedback from these references to converge towards an optimal prompt that maximizes performance across the task distribution. Specifically, COPLE consists of the following four parts:

Proxy Reference Tasks. To find the optimal $\boldsymbol{D}^*$ defined in (3), while avoiding data leakage, we first randomly sample a batch of reference tasks from the same distribution $\mathcal{D}$, denoted as $\mathcal{Z}_{ref}$. For example, when targeting the validation set of a dataset as downstream tasks, we construct the reference tasks by sampling from the training set. These reference tasks serve as a proxy for evaluating the prompt on the task distribution, which also accelerates COPLE, as evaluating on a small batch of examples is not as expensive as evaluating on the full validation set. Therefore, the optimal task description that COPLE tries to find can be transformed from (3) to:

$$\begin{aligned}\boldsymbol{D}^* &= \arg\min_{\boldsymbol{D}} \quad \mathcal{L}_{ref}(\boldsymbol{D}) \\ &= \arg\min_{\boldsymbol{D}} \mathop{\mathbb{E}}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{Z}_{ref}} [\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{P}_{\mathcal{Z}_{ref}}(\boldsymbol{X})), \boldsymbol{Y})]\end{aligned} \tag{4}$$

where $\boldsymbol{P}_{\mathcal{Z}_{ref}}$ denotes the entire prompt for the reference tasks $\mathcal{Z}_{ref}$.

Search by Word Influence. With the proxy reference tasks, COPLE then performs an iterative optimization process to find the optimal task description $\boldsymbol{D}^*$. As COPLE serves as a black-box method without accessing the gradient information of the model, we first define the influence of each word in the task description as the expected performance difference on proxy tasks when the word is deleted from the task description. Formally:

$$\mathcal{I}(d_i) = |\mathcal{L}_{ref}(\boldsymbol{D}) - \mathcal{L}_{ref}(\boldsymbol{D}_{\setminus i})| \tag{5}$$

where $\boldsymbol{D}_{\setminus i}$ denotes the task description with token d_i removed. For efficiency purposes, COPLE obtains the influence of each word only on the initial task description. Then, COPLE tries to iteratively find the optimal substitution for the most influential words in the descending order of their influence.

Lexical Search Space. To construct the search space $\mathcal{C}_i$ for token d_i in the task description, similar to that in §3.2, we reuse a pre-trained MLM to find semantically similar words. Formally, at each iteration t, we mask out the target d_i in current task description and feed the masked description into a pre-trained MLM f_{MLM}. The MLM then predicts a probability distribution over its vocabulary $\mathcal{V}$ for the masked position:

$$p(w|\boldsymbol{D}^{(t)}_{\setminus i}) = f_{MLM}(d_1, \ldots, \texttt{[MASK]}, \ldots, d_n) \tag{6}$$

where $w \in \mathcal{V}$, $\boldsymbol{D}^{(t)}_{\setminus i}$ denotes the task description at iteration t with token d_i masked out. We then select the top-k words with the highest probabilities and a empty token (delete) as the candidates.

Iterative Optimization. At each iteration t, COPLE selects the most influential token that has not been searched and constructs its corresponding search space according to (6). For each candidate $c \in \mathcal{C}_i$, COPLE substitutes d_i with c and evaluates the performance of the updated task description $\boldsymbol{D}^{(t)}$ on the small proxy reference tasks:

$$\mathcal{L}_{ref}(\boldsymbol{D}^{(t)}) = \mathop{\mathbb{E}}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{Z}_{ref}} [\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{P}^{(t)}_{\mathcal{Z}_{ref}}(\boldsymbol{X})), \boldsymbol{Y})] \tag{7}$$

where $\boldsymbol{D}^{(t)} = (d_1, \ldots, c, \ldots, d_n)$, and $\boldsymbol{P}^{(t)}_{\mathcal{Z}_{ref}}$ is the prompt with the task description $\boldsymbol{D}^{(t)}$ at iteration t. COPLE then selects the candidate c^* that minimizes the expected loss on the proxy reference tasks:

$$c^* = \arg\min_{c\in\mathcal{C}_i} \quad \mathcal{L}_{ref}(\boldsymbol{D}^{(t)}) \tag{8}$$

and updates the task description to $\boldsymbol{D}^{(t+1)}$ by replacing d_i with c^* if its performance is better than $\boldsymbol{D}^{(t)}$, i.e., if $\mathcal{L}_{ref}(\boldsymbol{D}^{(t+1)}) < \mathcal{L}_{ref}(\boldsymbol{D}^{(t)})$:

$$\boldsymbol{D}^{(t+1)} = (d_1, \ldots, c^*, \ldots, d_n) \tag{9}$$

otherwise, $\boldsymbol{D}^{(t+1)}$ is kept the same as $\boldsymbol{D}^{(t)}$. This process is repeated until all the most influential words are traversed (detailed in §4.1). Ideally, we take the found best $\boldsymbol{D}$ after optimization as the $\boldsymbol{D}^*$ defined in (4). The final optimized task description $\boldsymbol{D}^*$ is then used to construct the optimal prompt $\boldsymbol{P}^*_{\mathcal{Z}}(\boldsymbol{X})$ for the downstream tasks.

4 Experiment

4.1 Experiment Setup

Dataset and Model. We use GLUE (Wang et al., 2019) and MMLU (Hendrycks et al., 2021) for evaluation. For GLUE, we report the results on SST2 (Socher et al., 2013), CoLA (Warstadt et al., 2019), MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), RTE (Giampiccolo et al., 2007), MRPC (Dolan and Brockett, 2005), and QQP (Cer et al., 2017). For MMLU, we separately report the results on the subset of STEM, Humanities, Social Sciences, and Other. We use the `Llama-2-7B-chat` (Touvron et al., 2023), `Mistral-7B-Instruct-v0.1` (Jiang et al., 2023), and ChatGPT (`gpt-3.5-turbo-0125`) (OpenAI, 2022) as the target model. Please see Appendix A.1 for more details on datasets and models.

	GLUE							MMLU			
	SST2	CoLA	MNLI	QNLI	RTE	MRPC	QQP	STEM	Humanities	Soc.Sci	Other
`Llama-2-7B-chat`											
Original	90.71	33.27	35.51	51.57	53.07	68.06	27.58	28.04	23.94	35.31	38.59
w/ COPLE	**92.43**$_{0.70}$	**65.72**$_{1.29}$	**52.42**$_{1.33}$	**69.38**$_{2.10}$	**68.59**$_{4.08}$	**68.17**$_{0.53}$	**57.11**$_{1.80}$	**31.46**$_{0.54}$	**27.90**$_{1.77}$	**37.09**$_{0.79}$	**46.20**$_{0.40}$
`Mistral-7B-Instruct-v0.1`											
Original	87.27	74.31	65.18	75.22	64.98	54.41	68.92	35.19	37.07	52.23	51.83
w/ COPLE	**91.21**$_{1.48}$	**78.24**$_{1.58}$	**65.37**$_{0.89}$	**79.34**$_{0.35}$	**71.60**$_{0.55}$	**71.08**$_{0.98}$	**75.93**$_{0.15}$	**35.83**$_{0.54}$	**37.45**$_{0.51}$	**53.02**$_{0.17}$	**52.96**$_{0.28}$
ChatGPT (`gpt-3.5-turbo-0125`)											
Original	94.38	80.73	\	\	62.09	41.91	\	34.16	42.47	58.16	57.46
w/ COPLE	**94.80**$_{0.17}$	**82.91**$_{0.16}$	\	\	**80.35**$_{0.34}$	**70.75**$_{0.37}$	\	**36.45**$_{0.76}$	**43.44**$_{0.39}$	**58.46**$_{0.59}$	**57.61**$_{0.14}$

Table 1: Performance comparison (Accuracy) of models on GLUE and MMLU benchmarks using the human-crafted prompts (*Original*) with and without applying COPLE. The **bold** values indicate the better results, while the standard deviations are provided in smaller font. For MNLI, we report the average results on the *matched* and *mismatched* subsets. Some results for `gpt-3.5-turbo-0125` are denoted as "\", indicating that, due to the huge validation set and cost and efficiency considerations, corresponding experiments are not conducted.

Baseline. To show the effectiveness of COPLE and empirically demonstrate the conclusions in §3.2, we evaluate COPLE in the following scenarios: (i) *Original*: using human-crafted prompts from *HELM* (Lee et al., 2023) and *lm-evaluation-harness* (Gao et al., 2023). (ii) *In-context Learning*, following Brown et al. (2020), randomly concatenating 1 and 3 examples from the training set with *Original* manual prompts (as the $\boldsymbol{E}'$ in $\boldsymbol{P}_{\mathcal{Z}}(\boldsymbol{X})$), denoted as the *1-shot* and *3-shot* settings, respectively. (iii) *Emotion Prompt*: combining two different self-monitoring style emotional stimuli, used in (Li et al., 2023) with *Original* manual prompts, denoted as *EP02* and *EP03*, respectively. (iv) *Chain-of-thought*: combining zero-shot CoT trigger (Kojima et al., 2022) with *Original* manual prompts, denoted as *Zero-shot-CoT*. Please see Appendix A.3 for the detailed prompts used in evaluation.

Implementation Details. To construct the proxy reference tasks $\mathcal{Z}_{ref}$, we sample 100 tasks from training set. For the search space, we use RoBERTa (Liu et al., 2019) as the MLM in (6), selecting the top-30 highest probability substitutions for each iteration. Following (5), we take the 70% most influential words in a task description to perform optimization. We use HELM-style evaluation, with more details available in Appendix A.2. All reported average results and standard deviations are obtained from 3 runs with different seeds.

4.2 Main Results

Popular Prompts Suffer From Lexical Sensitivity. Table 1 shows the model performance on different tasks using *Original* human-crafted prompts and related prompts optimized by COPLE. The results demonstrate that even widely used human-crafted prompts fail to maximize model performance on downstream tasks, due to lexical sensitivity and specific words in prompts. Specifically, for `Llama-2-7B-chat`, the average accuracy across all datasets increased from 44.15% to 56.04% (11.89%↑) after applying COPLE. For `Mistral-7B-Instruct-v0.1`, the average accuracy increased from 60.60% to 64.73% (4.13%↑). ChatGPT also exhibited a notable improvement, with the average accuracy increasing from 58.92% to 65.59% (6.67%↑, 4 GLUE datasets + MMLU) when using prompts optimized by COPLE.

COPLE Recovers the Ability on Both Instruct-Following and Solving Downstream Tasks. Table 2 shows the model performance on various tasks using different prompts. Recall that the HELM-style evaluation used in our experiments may yield performance worse than random guessing, suggesting that models fail to complete the task as instructed by the provided prompt. Following this, we find that further modifications to the *Original* prompts, such as adding few-shot demo examples, may not always improve the performance. Such modifications may lead to the generation of incorrect or entirely irrelevant responses, such as repeating demo examples or generating non-existent new examples. Therefore, it can be deduced that the decline in model performance on downstream tasks may be attributed to a decreased ability of (i) problem-solving, as when the model gives wrong results, and (ii) instruct-following, as when the model gives irrelevant results. For example, for `Llama-2-7B-chat` on QQP, the *3-shot* accuracy decreases from 27.58% to 23.03% (4.55%↓)

	GLUE							MMLU			
	SST2	CoLA	MNLI	QNLI	RTE	MRPC	QQP	STEM	Humanities	Soc.Sci	Other
`Llama-2-7B-chat`											
In-Context Learning Prompts											
1-shot	57.68	68.74	19.81	49.57	51.62	68.14	3.48	26.17	21.43	25.82	22.98
w/ COPLE	**88.13**$_{3.16}$	**69.13**$_{0.29}$	**38.46**$_{5.36}$	**52.61**$_{0.62}$	**55.23**$_{0.51}$	**68.23**$_{0.12}$	**45.93**$_{17.25}$	**28.97**$_{0.31}$	**24.15**$_{0.27}$	**30.12**$_{0.63}$	**23.38**$_{0.23}$
3-shot	51.61	67.98	0.03	51.91	56.51	68.38	23.03	26.48	24.13	27.00	22.82
w/ COPLE	**70.07**$_{1.62}$	**68.36**$_{0.38}$	**31.17**$_{0.19}$	**55.78**$_{0.22}$	**57.76**$_{0.26}$	**68.59**$_{0.28}$	**57.61**$_{1.22}$	**30.22**$_{0.88}$	**26.38**$_{0.11}$	**29.82**$_{0.21}$	**25.45**$_{0.33}$
Emotion Prompts											
EP02	85.32	31.16	35.09	53.25	53.07	67.92	8.58	26.17	21.24	30.27	36.89
w/ COPLE	**92.26**$_{0.57}$	**67.98**$_{0.54}$	**50.15**$_{3.93}$	**64.59**$_{1.85}$	**57.16**$_{4.59}$	**68.42**$_{0.34}$	**30.55**$_{0.60}$	**31.78**$_{0.44}$	**28.09**$_{2.59}$	**35.91**$_{1.68}$	**40.14**$_{0.20}$
EP03	91.51	36.53	35.66	50.36	54.15	68.38	17.18	25.23	21.62	34.42	39.15
w/ COPLE	**92.78**$_{0.65}$	**68.41**$_{1.02}$	**52.23**$_{0.03}$	**67.40**$_{3.60}$	**62.09**$_{3.06}$	**68.63**$_{0.35}$	**42.34**$_{6.23}$	**33.02**$_{1.32}$	**27.12**$_{0.14}$	**36.65**$_{2.31}$	**43.80**$_{0.20}$
Chain-of-thought Prompts											
Zero-shot CoT	68.23	65.19	41.51	62.97	55.96	50.74	11.63	25.55	30.12	27.89	24.23
w/ COPLE	**84.29**$_{3.92}$	**67.45**$_{1.02}$	**50.51**$_{1.03}$	**69.64**$_{0.92}$	**59.81**$_{0.83}$	**66.42**$_{1.04}$	**23.49**$_{0.88}$	**27.73**$_{0.54}$	**32.72**$_{0.68}$	**34.42**$_{1.26}$	**31.27**$_{1.59}$
`Mistral-7B-Instruct-v0.1`											
In-Context Learning Prompts											
1-shot	81.08	34.13	58.87	51.55	51.99	31.86	43.64	28.66	31.66	42.14	45.92
w/ COPLE	**91.80**$_{0.41}$	**72.42**$_{0.77}$	**60.72**$_{1.46}$	**71.94**$_{0.18}$	**65.16**$_{2.81}$	**65.07**$_{0.52}$	**64.96**$_{0.16}$	**31.15**$_{0.44}$	**31.98**$_{0.73}$	**44.31**$_{0.95}$	**46.90**$_{0.28}$
3-shot	91.28	66.73	41.45	69.51	53.07	56.13	73.39	31.15	39.58	45.99	50.70
w/ COPLE	**93.43**$_{0.07}$	**72.77**$_{0.27}$	**56.35**$_{0.68}$	**75.62**$_{0.58}$	**68.77**$_{5.87}$	**67.65**$_{0.35}$	**75.42**$_{0.70}$	**35.31**$_{2.12}$	**39.67**$_{0.14}$	**49.11**$_{1.05}$	**51.97**$_{0.60}$
Emotion Prompts											
EP02	64.56	70.18	62.83	62.77	53.07	38.24	61.75	35.23	35.71	50.15	49.86
w/ COPLE	**89.11**$_{1.78}$	**75.36**$_{0.41}$	**64.36**$_{0.82}$	**72.57**$_{1.54}$	**72.56**$_{2.04}$	**70.10**$_{0.69}$	**72.43**$_{0.98}$	**37.07**$_{0.44}$	**37.64**$_{0.82}$	**52.08**$_{0.30}$	**51.97**$_{1.00}$
EP03	76.61	74.49	65.56	74.24	64.62	62.75	73.37	34.58	36.87	51.93	50.99
w/ COPLE	**91.44**$_{0.33}$	**76.32**$_{0.14}$	**67.24**$_{0.36}$	**76.37**$_{0.91}$	**72.38**$_{1.28}$	**74.39**$_{0.17}$	**75.10**$_{0.90}$	**36.76**$_{0.88}$	**37.45**$_{0.27}$	**52.67**$_{0.21}$	**52.02**$_{0.16}$
Chain-of-thought Prompts											
Zero-shot CoT	86.01	76.03	50.04	78.29	66.06	65.21	72.74	35.19	35.91	50.74	50.14
w/ COPLE	**90.90**$_{0.26}$	**76.27**$_{1.02}$	**67.41**$_{0.68}$	**79.45**$_{0.45}$	**71.48**$_{0.63}$	**74.14**$_{0.52}$	**76.80**$_{3.13}$	**36.34**$_{0.18}$	**36.87**$_{0.33}$	**52.08**$_{0.21}$	**52.54**$_{0.20}$

Table 2: Performance comparison (Accuracy) of models on GLUE and MMLU using different initial prompts with and without applying COPLE. The **bold** and smaller values denote better results and standard deviations.

compared to *Original*. However, without complex prompt engineering, minor lexical optimization performed by COPLE is enough to recover the declined ability, as the accuracy increases from 23.03% to 57.61% after applying COPLE, which also outperforms the original prompt optimized by COPLE (Table 1, 57.11%). Please see Appendix B.1 for the detailed optimized prompts.

			`Llama-2-7B-chat`			`Mistral-7B-Instruct-v0.1`		
		PPL.ori	U.Sim	BERTScore	PPL	U.Sim	BERTScore	PPL
GLUE	SST2	46	0.84	0.89	163	0.85	0.92	82
	CoLA	67	0.79	0.75	1875	0.81	0.87	412
	MNLI	635	0.75	0.82	805	0.78	0.88	1781
	QNLI	206	0.77	0.82	487	0.78	0.85	521
	RTE	635	0.74	0.78	2026	0.78	0.83	842
	MRPC	58	0.78	0.87	103	0.78	0.87	326
	QQP	185	0.72	0.76	656	0.78	0.86	306
MMLU	STEM	276	0.76	0.88	515	0.79	0.88	408
	Humanities		0.79	0.89	529	0.79	0.92	491
	Soc.Sci		0.80	0.91	494	0.80	0.94	479
	Other		0.78	0.84	663	0.79	0.87	525
	Avg.	267	0.77	0.84	756	0.79	0.88	561

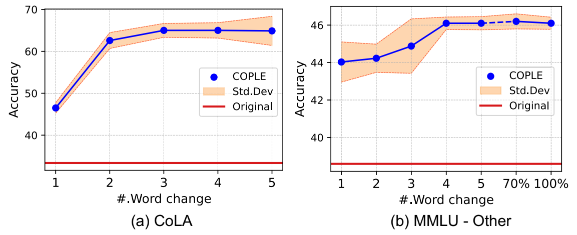
Table 3: Difference of semantic similarity and perplexity between original prompts and optimized prompts. *U.Sim* denotes the similarity obtained through USE.

4.3 Analysis and Ablation Study

In this section, we conduct further analysis and ablation studies on COPLE. When not specified, the results are obtained on `Llama-2-7B-chat`.

Difference Between Prompts. To measure the difference between original prompts and optimized prompts, we utilize Universal Sentence Encoder (USE) (Cer et al., 2018) and BERTScore (Zhang et al., 2020) to obtain semantic similarity. We also obtain their perplexity (PPL) (Jelinek et al., 1977) with GPT-2 (Radford et al., 2018b). Table 3 illustrates the differences between prompts. The USE similarity and BERTScore between the original and optimized prompts are consistently high across all tasks, indicating that the semantics of the prompts are well-preserved after optimization, which also confirms our conclusions in §3.2. However, the perplexity of the optimized prompts increases significantly compared to the original, indicating that using challenging words for the language model in prompts, rather than common words picked by humans, may help to improve the model performance on solving downstream tasks.

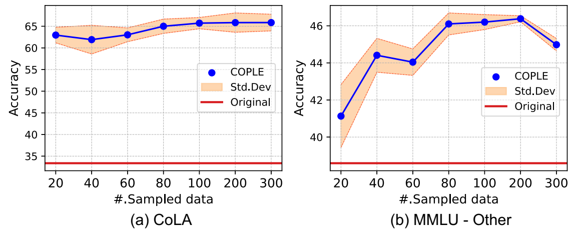
#.Word Change in Prompt. Figure 3 shows the impact of the number of words changed in prompt. Even if only changing the one word with the highest influence, the model performance signifi-



(a) CoLA

(b) MMLU - Other

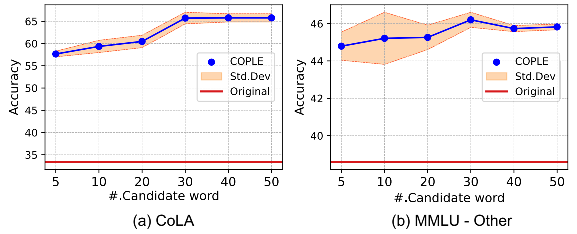
Figure 3: Impact of the number of words changed in prompt on downstream performance.



(a) CoLA

(b) MMLU - Other

Figure 4: Impact of the number of sampled examples in proxy reference tasks on downstream performance.



(a) CoLA

(b) MMLU - Other

Figure 5: Impact of the number of candidate words in search space on downstream performance.

cantly improves (CoLA: 33.27% to 46.50%$_{\pm 1.23\%}$, MMLU-Other: 38.59% to 44.04%$_{\pm 1.07\%}$). When the $\#$.*Word Change* increases, COPLE tends to achieve higher accuracy, while a moderate value is enough to achieve nearly the optimal result.

$\#$.Sampled Data for Proxy Tasks ($|\mathcal{Z}_{ref}|$). Figure 4 shows the impact of the size of proxy tasks. When proxy tasks contain a small number of 20 examples, COPLE still achieves notable improvements (CoLA: 33.27% to 62.96%$_{\pm 1.78\%}$; MMLU-Other: 38.59% to 41.13%$_{\pm 1.69\%}$). However, a larger size of sampled data helps find prompts with higher accuracy and lower standard deviations.

$\#$.Candidate Word in Search Space (k). Figure 5 shows the impact of the number of candidate words in search space. A small k is enough to support COPLE achieving considerable improvement (k=5, CoLA: 33.27% to 57.65$_{\pm 0.62\%}$, MMLU-Other: 38.59% to 44.79$_{\pm 0.75\%}$). Therefore, for efficient purposes, a small k is more suitable. However, using larger search space for each word is more likely to find prompts with better performance.

	STEM	Humanities	Social Sciences	Other
Word Influence	31.46$_{0.54}$	27.90$_{1.77}$	37.09$_{0.79}$	46.20$_{0.40}$
Random	30.32$_{0.88}$	25.48$_{1.64}$	36.20$_{1.29}$	44.69$_{1.14}$

Table 4: Ablation results on the search method. Random denotes searching on a random order of words.

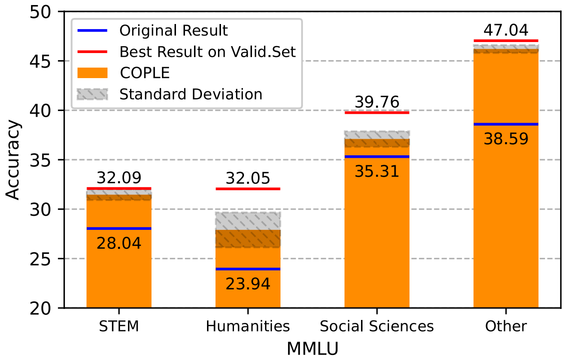


Figure 6: Performance difference when COPLE is performed on proxy reference tasks and on validation set.

	STEM	Humanities	Social Sciences	Other
Original	24.08	23.94	35.31	38.59
w/ COPLE	**31.46**	**27.90**	**37.09**	**46.20***
PromISe	27.10	24.51	32.64	40.28
w/ COPLE	**32.39***	**28.31***	**37.38***	**43.84**

Table 5: Performance comparison (Accuracy) of model on MMLU when combining COPLE with other prompt optimizing methods. Bold font indicates superior results when a method is combined with COPLE compared to without it, and * denotes the highest overall results.

Word Influence. Table 4 shows the ablation results of the search strategy related to word influence in COPLE. When replacing the search strategy with the random method, the average performance optimized by COPLE on MMLU decreases from 35.66% to 34.17%, and COPLE is less stable as the standard deviation gets larger.

How Far Is COPLE From the Optimal Results? Figure 6 shows the performance gap between the potential best prompts found by COPLE directly on the validation set (3) and on proxy tasks (4). The results show that the proxy tasks provide a reasonable approximation of the target task distribution, and the performance of COPLE is close to optimal.

COPLE Is Compatible With Other Prompt Optimization Methods. Recall that we find the prompts designed by humans always fails to maximize the performance of LLMs due to the lexical sensitivity. However, this lexical sensitivity could also impact the prompts that are further optimized by more complex prompt engineering methods. Table 5 shows the model performance when combining COPLE with other prompt optimization methods like PromISe (Wang et al., 2024). These results demonstrate that COPLE is not only compatible with other prompt optimization methods, but can also further improve model performance with lexical-only optimization. Therefore, we believe that lexical optimization should be a fundamental step, either before or after more complex prompt engineering methods, to maximize performance, and COPLE is an effective plug-and-play solution.

5 Conclusion

In this paper, we demonstrate the notable lexical sensitivity of LLMs to prompts, which potentially degrades their performance on downstream tasks. We show that even semantically similar prompts located in the neighborhood of the latent representation space may yield very different results. To recover the performance drop caused by the sensitivity, we propose COPLE, a black-box combinatorial optimization framework that iteratively improves lexical choices in prompts. Experiments illustrate the effectiveness of COPLE in recovering both the model's ability of instruct-following and solving downstream tasks. We believe that carefully checking the word usage is essential before performing complex prompt engineering.

Acknowledgements

This work was supported by the Youth Innovation Promotion Association CAS (No.2023166), and the Informatization Plan of Chinese Academy of Sciences (Grant No. CAS-WX2021SF-0508).

Limitations

Despite the effectiveness of COPLE, we want to discuss some limitations of this work. Firstly, our experimental scope is primarily restricted to models around the 7-billion-parameter scale, as our computational resources are limited. Secondly, while we focus on optimizing the lexical choices within the task description component of the prompts, it is possible that lexical sensitivity affects the entirety of a prompt. However, expanding our optimization to include the full prompt significantly increases the size of search space, making the experiment computationally infeasible with our current resources. Despite these limitations, our study provides insights into the influence of lexical variation on language model prompts, from both the perspective of downstream performance and latent sentence representation. The findings highlight that even subtle lexical changes can significantly enhance the performance of language models on downstream tasks.

Ethics Statement

In this paper, we highlight the lexical sensitivity of LLMs, which can be susceptible to exploitation for malicious purposes. Although our primary objective is to enhance the understanding of LLMs and improve their performance in beneficial applications, we acknowledge the dual-use risk inherent in revealing this property. Malicious actors might exploit our findings to craft prompts that increase the persuasiveness of disinformation or other harmful content. Nonetheless, we believe that our work will ultimately contribute to a deeper understanding of the capabilities and limitations of LLMs, thereby facilitating the development of more robust and secure language models. We employ publicly available datasets that exclude sensitive or personally identifiable information (PII) and comply with their respective licenses. Our research adheres to strict ethical guidelines, demonstrating our methods in a manner that avoids unintended harm.

References

Alice F. Healy Adam Drewnowski. 1978. Detection errors on the word the: Evidence for the acquisition of reading levels. *Memory & Cognition*, 5.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6. Venice.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *Tac*.

Charles E. Blair. 1990. Integer and combinatorial optimization (george l. nemhauser and laurence a. wolsey). *SIAM Rev.*, 32(2).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. Ieee.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint*, abs/1803.11175.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, I~nigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*.

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, NeurIPS*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL@ACL 2007 Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *CoRR*, abs/2307.11760.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. *CoRR*, abs/1907.11692.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *CoRR*, abs/2209.07686.

Leo X McCusker, Philip B Gough, and Randolph G Bias. 1981. Word recognition inside out and outside in. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3).

OpenAI. 2022. Introducing chatgpt. In *OpenAI Blog*.

Long Ouyang, Jeffrey 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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. Grips: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training. In *OpenAI Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Keith Rayner, Sarah J. White, Rebecca L. Johnson, and Simon P. Liversedge. 2006. Raeding wrods with jubmled lettres: There is a cost. *Psychological Science*, 17(3). Pmid: 16507057.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86).

Guy C. van Orden. 1987. A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, 15(3).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR*.

Minzheng Wang, Nan Xu, Jiahao Zhao, Yin Luo, and Wenji Mao. 2024. PromISe: Releasing the capabilities of LLMs with prompt introspective search. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *CoRR*, abs/2305.14688.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *CoRR*, abs/2309.03409.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pengwei Zhan, Yang Wu, Shaolei Zhou, Yunjian Zhang, and Liming Wang. 2022a. Mitigating the inconsistency between word saliency and model confidence with pathological contrastive training. In *Findings of the Association for Computational Linguistics: ACL*.

Pengwei Zhan, Jing Yang, Xiao Huang, Chunlei Jing, Jingying Li, and Liming Wang. 2023a. Contrastive learning with adversarial examples for alleviating pathology of language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Pengwei Zhan, Jing Yang, He Wang, Chao Zheng, Xiao Huang, and Liming Wang. 2023b. Similarizing the influence of words with contrastive learning to defend word-level adversarial text attack. In *Findings of the Association for Computational Linguistics: ACL*.

Pengwei Zhan, Jing Yang, He Wang, Chao Zheng, and Liming Wang. 2024. Rethinking word-level adversarial attack: The trade-off between efficiency, effectiveness, and imperceptibility. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Pengwei Zhan, Chao Zheng, Jing Yang, Yuxiang Wang, Liming Wang, Yang Wu, and Yunjian Zhang. 2022b. PARSE: an efficient search method for black-box adversarial text attacks. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING*.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E. Gonzalez. 2022. TEMPERA: test-time prompting via reinforcement learning. *CoRR*, abs/2211.11890.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR*.

A Additional Experimental Details

A.1 Details on Dataset

The General Language Understanding Evaluation (GLUE) (Wang et al., 2019) benchmark is a collection of datasets for training, evaluating, and analyzing natural language understanding systems. The subset used in our experiment include: (1) The Stanford Sentiment Treebank (SST2) (Socher et al., 2013) consists of movie review sentences annotated for sentiment. (2) The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) contains English acceptability judgments drawn from linguistic theory publications. (3) The Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018) corpus includes sentence pairs annotated with textual entailment information. (4) The Question-answering NLI (QNLI) (Rajpurkar et al., 2016) is derived from SQuAD, converted to a binary sentence pair classification task. (5) The Recognizing Textual Entailment (RTE) datasets come from a series of textual entailment challenges (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009). (6) The Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) contains sentence pairs annotated for semantic equivalence. (7) The Quora Question Pairs2 (QQP) (Cer et al., 2017) includes question pairs from Quora annotated for semantic equivalence. The Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) contains multiple-choice questions that cover 57 tasks, which can be divided into four main subsets: STEM, Humanities, Social Sciences, and Other, with 14,042 test and 1,531 validation examples. In our experiment, for constructing the proxy reference tasks, we sample from the test set on MMLU. More information about the datasets is provided in Table 6.

Dataset	#.Training example	#.Validation example	Mission Type	#.Category
SST2	67 349	872	Sentiment Analysis	2
CoLA	8551	1043	Linguistic Acceptability	2
MNLI	392 702	19 647	Natural Language Inference	3
QNLI	104 743	5463	Natural Language Inference	2
RTE	2490	277	Natural Language Inference	2
MRPC	3668	408	Semantic Equivalence	2
QQP	363 849	40 430	Semantic Equivalence	2
MMLU	14042 (Test)	1531	Question Answering	4

Table 6: Summary of datasets used in the experiments. For MNLI, 19,647 validation examples consists of 9,815 from the matched in-domain section and 9,832 from the mismatched cross-domain section. For MMLU, we report the size of test set rather than training set.

A.2 Details on Evaluation

We follow the same evaluation style as HELM (Lee et al., 2023), which expects the model to output the correct label word (for GLUE) or the letter of the correct option (for MMLU), rather than directly using the probability of the output token for judgement. For example, on SST2, for a sentence with positive sentiment, we expect the model output to be "*positive*". Similarly, on MMLU, the model is expected to output one of the letters "*A*", "*B*", "*C*", or "*D*" that matches the correct answer. Otherwise, we consider the model makes incorrect decisions. Note that this evaluation approach may result in model performance worse than that of random guessing. However, we believe that it provides a more accurate indication of the model's ability on instruction following and on solving downstream tasks. For all datasets, we report the performance with Accuracy.

A.3 Details on Baseline Prompts

In the main text, we perform COPLE on various scenarios, including *Original*, *1-shot*, *3-shot*, *EP02*, *EP03*, and *Zero-shot-CoT*; here we report the detailed prompts of these scenarios. The initial prompts for GLUE are in Table 7-13, and the initial prompts for MMLU are in Table 14-15. It should also be noted that for scenarios of *Emotion Prompts*, following Li et al. (2023), we insert the additional text at the end of the task description and verbalizer, but before the demo examples, and we do not consider the additional text to be a part of the task descriptions to perform optimization.

Dataset	Scenario	Prompt
SST2	Original	For the given sentence, label the sentiment of the sentence as positive or negative. Do not respond with anything other than the labels ‘positive’ or ‘negative’. Question: {content} Answer:

	1-shot	For the given sentence, label the sentiment of the sentence as positive or negative. Do not respond with anything other than the labels ‘positive’ or ‘negative’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	For the given sentence, label the sentiment of the sentence as positive or negative. Do not respond with anything other than the labels ‘positive’ or ‘negative’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	For the given sentence, label the sentiment of the sentence as positive or negative. Do not respond with anything other than the labels ‘positive’ or ‘negative’. This is very important to my career. Question: {content} Answer:
	EP03	For the given sentence, label the sentiment of the sentence as positive or negative. Do not respond with anything other than the labels ‘positive’ or ‘negative’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	For the given sentence, label the sentiment of the sentence as positive or negative. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘positive’ / ‘negative’</label>." Question: {content} Answer:

Table 7: Detailed baseline prompts for SST2. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	**Scenario**	**Prompt**
CoLA	Original	Does this sentence make sense? Do not respond with anything other than the labels ‘Yes’ or ‘No’. Question: {content} Answer:

	1-shot	Does this sentence make sense? Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Does this sentence make sense? Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Does this sentence make sense? Do not respond with anything other than the labels ‘Yes’ or ‘No’. This is very important to my career. Question: {content} Answer:
	EP03	Does this sentence make sense? Do not respond with anything other than the labels ‘Yes’ or ‘No’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Does this sentence make sense? Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘Yes’ / ‘No’</label>." Question: {content} Answer:

Table 8: Detailed baseline prompts for CoLA. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt
MNLI	Original	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘entailment’, ‘neutral’, or ‘contradiction’. Question: {content} Answer:

	1-shot	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘entailment’, ‘neutral’, or ‘contradiction’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘entailment’, ‘neutral’, or ‘contradiction’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘entailment’, ‘neutral’, or ‘contradiction’. This is very important to my career. Question: {content} Answer:
	EP03	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘entailment’, ‘neutral’, or ‘contradiction’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Please identify whether the premise entails the hypothesis. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘entailment’ / ‘neutral’ / ‘contradiction’</label>. " Question: {content} Answer:

Table 9: Detailed baseline prompts for MNLI. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	**Scenario**	**Prompt**
QNLI	Original	Please identify whether the sentence answers the question. Do not respond with anything other than the labels ‘Yes’ or ‘No’. Question: {content} Answer:

	1-shot	Please identify whether the sentence answers the question. Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Please identify whether the sentence answers the question. Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Please identify whether the sentence answers the question. Do not respond with anything other than the labels ‘Yes’ or ‘No’. This is very important to my career. Question: {content} Answer:
	EP03	Please identify whether the sentence answers the question. Do not respond with anything other than the labels ‘Yes’ or ‘No’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Please identify whether the sentence answers the question. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘Yes’ / ‘No’</label>." Question: {content} Answer:

Table 10: Detailed baseline prompts for QNLI. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt
RTE	Original	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘Yes’ or ‘No’. Question: {content} Answer:

	1-shot	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘Yes’ or ‘No’. This is very important to my career. Question: {content} Answer:
	EP03	Please identify whether the premise entails the hypothesis. Do not respond with anything other than the labels ‘Yes’ or ‘No’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Please identify whether the premise entails the hypothesis. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘Yes’ / ‘No’</label>." Question: {content} Answer:

Table 11: Detailed baseline prompts for RTE. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt
MRPC	Original	Do both sentences mean the same thing? Do not respond with anything other than the labels ‘Yes’ or ‘No’. Question: {content} Answer:

	1-shot	Do both sentences mean the same thing? Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Do both sentences mean the same thing? Do not respond with anything other than the labels ‘Yes’ or ‘No’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Do both sentences mean the same thing? Do not respond with anything other than the labels ‘Yes’ or ‘No’. This is very important to my career. Question: {content} Answer:
	EP03	Do both sentences mean the same thing? Do not respond with anything other than the labels ‘Yes’ or ‘No’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Do both sentences mean the same thing? Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘Yes’ / ‘No’</label>." Question: {content} Answer:

Table 12: Detailed baseline prompts for MRPC. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt
QQP	Original	Please identify whether the sentences have the same meaning. Do not respond with anything other than the labels ‘equal’ or ‘unequal’. Question: {content} Answer:

	1-shot	Please identify whether the sentences have the same meaning. Do not respond with anything other than the labels ‘equal’ or ‘unequal’.
		
		Question: {demo_content}
		Answer: {demo_answer}
		
		Question: {content}
		Answer:

	3-shot	Please identify whether the sentences have the same meaning. Do not respond with anything other than the labels ‘equal’ or ‘unequal’.
		
		Question: {demo_content_1}
		Answer: {demo_answer_1}
		
		Question: {demo_content_2}
		Answer: {demo_answer_2}
		
		Question: {demo_content_3}
		Answer: {demo_answer_3}
		
		Question: {content}
		Answer:

	EP02	Please identify whether the sentences have the same meaning. Do not respond with anything other than the labels ‘equal’ or ‘unequal’. This is very important to my career. Question: {content} Answer:
	EP03	Please identify whether the sentences have the same meaning. Do not respond with anything other than the labels ‘equal’ or ‘unequal’. You’d better be sure. Question: {content} Answer:
	Zero-shot CoT	Please identify whether the sentences have the same meaning. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘equal’ / ‘unequal’</label>." Question: {content} Answer:

Table 13: Detailed baseline prompts for QQP. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt

MMLU	Original	The following are multiple choice questions (with answers) about {task}. Do not respond with anything other than the answer labels ‘A’, ‘B’, ‘C’, or ‘D’.
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

	1-shot	The following are multiple choice questions (with answers) about {task}. Do not respond with anything other than the answer labels ‘A’, ‘B’, ‘C’, or ‘D’.
		
		Question: {demo_question}
		A. {demo_option_A}
		B. {demo_option_B}
		C. {demo_option_C}
		D. {demo_option_D}
		
		Answer: {demo_answer}
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

	3-shot	The following are multiple choice questions (with answers) about {task}. Do not respond with anything other than the answer labels ‘A’, ‘B’, ‘C’, or ‘D’.
		
		Question: {demo_question_1}
		A. {demo_option_A_1}
		B. {demo_option_B_1}
		C. {demo_option_C_1}
		D. {demo_option_D_1}
		
		Answer: {demo_answer_1}
		
		Question: {demo_question_2}
		A. {demo_option_A_2}
		B. {demo_option_B_2}
		C. {demo_option_C_2}
		D. {demo_option_D_2}
		
		Answer: {demo_answer_2}
		
		Question: {demo_question_3}
		A. {demo_option_A_3}
		B. {demo_option_B_3}
		C. {demo_option_C_3}
		D. {demo_option_D_3}
		
		Answer: {demo_answer_3}
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

Table 14: Detailed baseline prompts of *Original* and *In-context Learning* setting for MMLU. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

Dataset	Scenario	Prompt

MMLU	EP02	The following are multiple choice questions (with answers) about {task}. Do not respond with anything other than the answer labels ‘A’, ‘B’, ‘C’, or ‘D’. This is very important to my career.
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

	EP03	The following are multiple choice questions (with answers) about {task}. Do not respond with anything other than the answer labels ‘A’, ‘B’, ‘C’, or ‘D’. You’d better be sure.
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

	Zero-shot CoT	The following are multiple choice questions (with answers) about {task}. Let’s think step by step. Then, end the response with "Therefore, the answer is: <label>‘A’ / ‘B’ / ‘C’ / ‘D’</label>."
		
		Question: {question}
		A. {option_A}
		B. {option_B}
		C. {option_C}
		D. {option_D}
		
		Answer:

Table 15: Detailed baseline prompts of *Emotion Prompt* and *Chain-of-thought* setting for MMLU. In the prompt, brown text indicates the task description, which is the target that COPLE performs on, green text indicates the verbalizer, and blue text indicates the demo examples for few-shot learning. “*{text}*” denotes the placeholder, which will be replaced with the text sampled from the dataset.

B Additional Experimental Results

B.1 Details on the Optimized Prompt Crafted by COPLE

Table 16-17 shows the optimal task descriptions optimized by COPLE on `Llama2-7B-chat` for various scenarios. Table 18-19 shows the optimal task descriptions optimized by COPLE on `Mistral-7B-Instruct-v0.1` for various scenarios. Table 20 shows the optimal task description optimized by COPLE on ChatGPT (`gpt-3.5-turbo-0125`) and the *Original* scenario.

Dataset	Scenario	Task Description (Before)	Task Description (After)
SST2	Original	For the given sentence, label the sentiment of the sentence as positive or negative.	Before the given sentence, label the sentiment of the user as positively or negative.
	1-shot		For the closing sentence, label this rest of next sentence as positive or positive.
	3-shot		For another given sentence, label the comment preceding no sentence as angry or negative.
	EP02		Regarding ANY OTHER NEWS, label the sentiment of the sentence positive or negative.
	EP03		If the given sentence, label the quality of the statement as positively or negative.
	Zero-shot CoT		For the introductory discussion, label the sentiment of the reviewer for positive or opposite.
CoLA	Original	Does this sentence make sense?	Did THIS sentence make?
	1-shot		Does every word make Sense?
	3-shot		Does my sentence any sense?
	EP02		Has this now make sense?
	EP03		Made THIS sentence make it?
	Zero-shot CoT		Should THAT sentence made sense?
MNLI	Original	Please identify whether the premise entails the hypothesis.	Please indicate whether the content entails valid conclusion.
	1-shot		Please identify all the elements entails the hypothesis.
	3-shot		Please identify whether the premise is the facts.
	EP02		must identify whether the premise opposes the hypothesis.
	EP03		Results evaluate Whether your data entails the hypothesis.
	Zero-shot CoT		Note whether the answer satisfies the evidence.
QNLI	Original	Please identify whether the sentence answers the question.	Please specify whether the above answers the question.
	1-shot		Please comment whether sentence answers the question.
	3-shot		We consider whether the sentence answers the question.
	EP02		Please assess whether the above answered the error.
	EP03		Please correct whether this answered the question.
	Zero-shot CoT		Please assess whether the sentence answers the queries.
RTE	Original	Please identify whether the premise entails the hypothesis.	First estimate whether premise entails each hypothesis.
	1-shot		Please confirm whether current knowledge advances the hypothesis.
	3-shot		Now identify whether the correction entails the hypothesis.
	EP02		First identify whether the premise confirms each hypothesis.
	EP03		Please state whether the conclusion justifies the hypothesis.
	Zero-shot CoT		Please identify whether certain premise entails correct hypothesis.
MRPC	Original	Do both sentences mean the same thing?	Do both sentences mean a same thing?
	1-shot		Do both languages mean the same thing?
	3-shot		Do both letters mean the same thing?
	EP02		Do parallel sentences repeat the same thing?
	EP03		Are both sentences mean the same thing?
	Zero-shot CoT		did both sentences convey the same thing?
QQP	Original	Please identify whether the sentences have the same meaning.	Please identify since the posts have the same text.
	1-shot		will identify Where unrelated sentences have nearly same significance.
	3-shot		Please identify whether the sentences have the same ending.
	EP02		Please determine are the have the same meaning.
	EP03		Please identify both your works have the same subject.
	Zero-shot CoT		Please note whether adjacent sentences display the identical meaning.

Table 16: Detailed optimized task descriptions on `Llama2-7B-chat` and GLUE.

Dataset	Scenario	Task Description (Before)	Task Description (After)
MMLU-STEM	Original	The following are multiple choice questions (with answers) about {task}.	THE exercises contain easy choice questions (with answers) about {task}.
	1-shot		The following generates multiple discussion questions (using arrows) about {task}.
	3-shot		Se following are multiple popular questions (with answers) about {task}.
	EP02		The examples list multiple choice questions (with labels) without {task}.
	EP03		The are infinite choice questions (by data) about {task}.
	Zero-shot CoT		The following are multiple choice problems (default answers) involving {task}.
MMLU-Humanities	Original		The model identifies multiple popular questions (which answers) about {task}
	1-shot		In following are all survey questions (with answers) about {task}.
	3-shot		The following are multiple questions (with answers) named {task}.
	EP02		Graph following illustrates third choice answer (simple answers) about {task}.
	EP03		The candidates are multiple choice questions (with) about {task}.
	Zero-shot CoT		The follows are most popular questions (mostly answers) about {task}.
MMLU-Social Sciences	Original		The following displays multiple choice questions (with answers) about {task}.
	1-shot		The following multiple popular questions (with answers) about {task}.
	3-shot		The following are multiple different questions (with answers) about {task}.
	EP02		The following summarizes the choice questions (complete ones) about {task}.
	EP03		The entries present multiple choice questions (with answers) involving {task}.
	Zero-shot CoT		following are some choice questions (complete answers) about {task}.
MMLU-Other	Original		The candidates posted multiple choice questions (with answers) about {task}.
	1-shot		The following are choice questions (with answers) about {task}.
	3-shot		The following are multiple standard answers (with answers) about {task}.
	EP02		The Following are the choice questions (and answers) following {task}.
	EP03		The following implements multiple choice questions (No exceptions) about {task}.
	Zero-shot CoT		The following contains your first questions (with explanation) about {task}.

Table 17: Detailed optimized task descriptions on `Llama2-7B-chat` and MMLU. "*{task}*" denotes the placeholder, which will be replaced with the detailed subset type.

Dataset	Scenario	Task Description (Before)	Task Description (After)
SST2	Original	For the given sentence, label the sentiment of the sentence as positive or negative.	For the given sentence, label the sentiment of the sentence as positive or constructive.
	1-shot		For the given sentence, keep positive sentiment of the sentence as positive not negative.
	3-shot		Before the given sentence, express components of the sentence as positive OR negatively.
	EP02		After the desired context, label every result of the process as happy or negative.
	EP03		For my given question, label the sentiment of the sentence as positive or optimistic.
	Zero-shot CoT		For the given expression, label the outcome of the sentence as positive or.
CoLA	Original	Does this sentence make sense?	Does this sentence form follows?
	1-shot		Does each word contain sense?
	3-shot		Does each sentence make points?
	EP02		Do I sentence make sense?
	EP03		Does each sentence make sense?
	Zero-shot CoT		Has this sentence make sense?
MNLI	Original	Please identify whether the premise entails the hypothesis.	Please assess how the result entails proposed hypothesis.
	1-shot		Please assess whether sufficient premise entails the claim.
	3-shot		Please identify between and premise the hypothesis.
	EP02		Please show whether the answer entails the hypothesis.
	EP03		Please identify whether the result entails the hypothesis.
	Zero-shot CoT		Please evaluate whether the hypothesis entails my observations.
QNLI	Original	Please identify whether the sentence answers the question.	Please identify whether any sentence asked either question.
	1-shot		Please repeat in the affirmative regarding the question.
	3-shot		please identify unless our article answers further question.
	EP02		must identify whether the article answers the question.
	EP03		Please identify whether the sentence addressed previous question.
	Zero-shot CoT		Please identify whether her sentence supports the question.
RTE	Original	Please identify whether the premise entails the hypothesis.	Please find sure the premise matches any hypothesis.
	1-shot		Please assess whether the premise supports the premises.
	3-shot		Please identify HOW either premise fits any other.
	EP02		Please verify either possible facts entails the claims.
	EP03		Please verify when the claim matches the fact.
	Zero-shot CoT		Please identify that The premise entails the hypothesis.
MRPC	Original	Do both sentences mean the same thing?	Do both sentences suggest the same picture?
	1-shot		Do other sentences suggest a same theme?
	3-shot		Do other sources suggest the Same thing?
	EP02		both sentences mean the whole thing?
	EP03		Do both sentences explain the same thing?
	Zero-shot CoT		Do both sentences describe the whole thing?
QQP	Original	Please identify whether the sentences have the same meaning.	Please identify whether following articles have the same keywords.
	1-shot		Please compare whether these sentences match the exact meaning.
	3-shot		Please identify whenever individual sentences have the equivalent content.
	EP02		Please tell whether two quotes have the identical message.
	EP03		Please identify whether the arguments have the same context.
	Zero-shot CoT		Please identify between the sentences have the same value.

Table 18: Detailed optimized task descriptions on `Mistral-7B-Instruct-v0.1` and GLUE.

Dataset	Scenario	Task Description (Before)	Task Description (After)
MMLU-STEM	Original	The following are multiple choice questions (with answers) about {task}.	The following are multiple choice questions (their variants) about {task}.
	1-shot		the following two first choice questions (with answers) about {task}.
	3-shot		Examples Here were multiple detailed questions (full comments) about {task}.
	EP02		The mes are multiple choice questions (with parentheses) about {task}.
	EP03		Examples following are multiple example questions (complete answers) about {task}.
	Zero-shot CoT		Then following are binary logic questions (with tags) about {task}.
MMLU-Humanities	Original		The are three choice cases (with answers) about {task}.
	1-shot		The Following are multiple choice questions (with hints) about {task}.
	3-shot		Then following are multiple choice questions (with answers) about {task}.
	EP02		The following are multiple obvious choices (easy fixes) using {task}.
	EP03		The Following are multiple choice messages (with answers) about {task}.
	Zero-shot CoT		The Below are multiple hypothetical questions (with answers) about {task}.
MMLU-Social Sciences	Original		The following are choice questions (with answers) about {task}.
	1-shot		The following are ten sample questions (with keywords) about {task}.
	3-shot		The following are two sample questions (with answers) about {task}.
	EP02		The Following are multiple choice questions (with solutions) containing {task}.
	EP03		The Following are one choice questions (with answers) about {task}
	Zero-shot CoT		The followed five multiple choice Questions (with Answers) about {task}.
MMLU-Other	Original		The following are multiple choice questions (answers) and {task}.
	1-shot		Then follow are two choice questions (with answers) about {task}.
	3-shot		Both following are multiple boolean questions (with answers) about {task}.
	EP02		The answers are Multiple choice questions (with answers) and {task}.
	EP03		The following are multiple nested questions (with answers) about {task}.
	Zero-shot CoT		Ch following is multiple choice questions (with answers) about {task}.

Table 19: Detailed optimized task descriptions on `Mistral-7B-Instruct-v0.1` and MMLU. "*{task}*" denotes the placeholder, which will be replaced with the detailed subset type.

Dataset	**Task Description (Before)**	**Task Description (After)**
SST2	For the given sentence, label the sentiment of the sentence as positive or negative.	For the given article, label the sentiment above the sentence as positive or negative.
CoLA	Does this sentence make sense?	this sentence make sense?
RTE	Please identify whether the premise entails the hypothesis.	Please evidence the premise support current hypothesis.
MRPC	Do both sentences mean the same thing?	Do both answers mean this correct thing?
MMLU-STEM	The following are multiple choice questions (with answers) about {task}.	The following are infinite choice questions (NO answers) within {task}.
MMLU-Humanities		Items above are a choice (with answers) about {task}.
MMLU-Social Sciences		The following answers multiple choice question (with answers) about {task}.
MMLU-Other		The Following answers multiple choice questions (with answers) about {task}.

Table 20: Detailed optimized task descriptions on ChatGPT (`gpt-3.5-turbo-0125`) on the *Original* scenario. "*{task}*" denotes the placeholder, which will be replaced with the detailed subset type.