Do prompt positions really matter?

Anonymous ACL submission

Abstract

Prompt-based models have gathered a lot of attention from researchers due to their remarkable advancements in the fields of zero-shot and few-shot learning. Developing an effective prompt template plays a critical role. However, prior studies have mainly focused on prompt vocabulary searching or embedding initialization within a predefined template with the prompt position fixed. In this empirical study, we conduct the most comprehensive analysis to date of prompt position for diverse Natural Language Processing (NLP) tasks. Our findings quantify the substantial impact prompt position has on model performance. We observe that the prompt position used in prior studies is often sub-optimal, and this observation is consistent even in widely used instruction-tuned models. These findings suggest prompt position optimisation as a valuable research direction to augment prompt engineering methodologies and prompt position-aware instruction tuning as a way to build more robust models in the future.

1 Introduction

Recently, Brown et al. (2020) have shown the impressive performance of using handcrafted prompts with a frozen language model in zero-shot and few-shot learning, leading to of interest and increased activity in prompt engineering within the NLP community (Schick and Schütze, 2020; Gao et al., 2020; Li and Liang, 2021). Prompting (a.k.a prompt-based learning (Liu et al., 2021a)) aims to reformat an NLP problem to match as closely as possible the format used in the original training task of the pre-trained language model used. To apply prompt-based learning method effectively, a critical step involves the creation of a prompt template that maximizes performance on the downstream task.

In many previous works, it is common to manually pre-define a template while keeping the prompt position fixed (e.g. prepend the prompt to the input (Lester et al., 2021)). These studies often concentrate more on either prompt vocabulary searching (Gao et al., 2020; Shin et al., 2020; Ben-David et al., 2021) or prompt embedding initialization (Liu et al., 2021c; Gu et al., 2022). However, there has been limited research exploring how different approaches to positioning the prompt sequences can affect the models' behaviour, despite indications that varying prompt positions may lead to performance difference (Mao et al., 2022; Wu et al., 2022).

Hence, in this paper, we quantify how much prompt position matters by evaluating various accessible models on different NLP tasks under few-shot and zero-shot settings. We comprehensively test a range of prompt position options with many widely used prompt styles (e.g. cloze and prefix) and methods (e.g. gradient-based and gradient-free). Our findings reveal unexpected performance variations among different prompt positions in both zero-shot and few-shot settings. We also discover that instruction-tuned models do not always reduce performance disparities, even though they typically include vocabulary and positional variations in their training templates. Interestingly, we observe that in many cases, the prompt positions used in previously published work show a sub-optimal performance compared to other prompt position choices. Our choice of zero and few-shot tasks were driven by the observation that prompting methods are particularly useful when training data is limited (Liu et al., 2021a), and this hypothesis is born out by our results which show prompt position matters more when labelled datasets are smaller.

The key contributions of this paper are[1]:

- To the best of our knowledge, we are the first comprehensive analysis looking at the impact of prompt position across different methods

[1]Our model code and checkpoints are available to allow full reproducibility of this work `URI.if.paper.accepted`.

and prompt types in both few-shot and zero-shot settings for a variety of NLP tasks.

- Empirical results show that the prompt positions matter. The positions used in many published works are often sub-optimal choices, with no universally superior prompt position across all tasks. These results suggest prompt position optimisation might be a useful addition to the existing field of prompt engineering and prompt position-aware instruction tuning could be explored to build more robust models in the future.

2 Related Work

Prompt-based learning. Many prior works have concentrated on **Gradient-based** methods within discrete spaces (Schick and Schütze, 2020; Schick and Schütze, 2020; Gao et al., 2020; Shin et al., 2020) as well as prompting directly in the embedding space. This latter approach uses tunable prompt tokens that are not limited to natural language, which can be either prepended to the input (Lester et al., 2021; Liu et al., 2021b; Gu et al., 2022) or be inserted in a hybrid template (Liu et al., 2021c). Sun et al. (2022) optimize continuous tokens without using gradients. However, this approach is not suitable for APIs like GPT-3, which only allows text modifications rather than token embeddings. There are also **Gradient-free** works focusing on in-context learning (Brown et al., 2020; Lu et al., 2021), chain-of-thought (Wei et al., 2022b; Yao et al., 2023), and instruction generation (Prasad et al., 2022; Zhou et al., 2022) especially when instruction tuning (Sanh et al., 2022; Wei et al., 2022a; Chung et al., 2022) plays a key role in the steering process of Large Pre-trained LMs. Our paper includes experiments examining prompt positions from both gradient-based and gradient-free perspectives.

Prompt position. There is limited work that involves prompt positions. Mao et al. (2022) find that the position of a handcrafted prompt (before or after the input) affects model performance, but there's no consensus on which position is best. Wu et al. (2022) propose an instance-dependent prompt generation method; meanwhile, they study the effect of inserting a sequence of prompt tokens in different positions based on their proposed method and prompt tuning (Lester et al., 2021). Recently, Yang et al. (2023) have proposed a dynamic position method that can significantly improve the performance of prompt tuning. They both point out that different positions of prompts will deliver different results with the consideration of only one specific approach to creating the prompt. In this paper, we present the most comprehensive analysis of prompt positions to date and take into account various types of prompts under both zero-shot and few-shot settings.

3 Method

3.1 Prompt Style

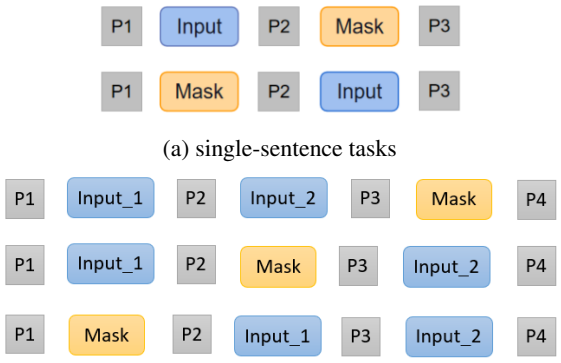
Two common styles of prompts are explored in our experiments: **Cloze style** aims to let LMs fill in the blanks. For example, the input of sentiment classification *"I love this movie"* can be formulated as *"I love this movie. Overall, it was a [Mask] movie."*, and the model will be asked to predict the masked token. **Prefix style** aims to let LMs generate an answer given a prefix, which means the entire input comes before the final prediction. For example, the input *"I love this movie"* will be formulated into *"I love this movie. Is this review positive or negative?"*, and the model will be asked to generate the answer.

3.2 Prompt Position

Prompt position is the variable of interest in our study. We take into account the position where prompt tokens can be inserted and enumerate a broad range of permutations to test.

Concretely, for **Cloze style** prompts, as shown in Figure 1, we consider the relative position of the [mask] token to the input. There are m types of Input-[MASK] concatenations ($m = 2$ for single-sentence and $m = 3$ for sentence-pair tasks[2]), each with n potential locations that could insert prompt sequences ($n = 3$ and $n = 4$, respectively). In contrast to Wu et al. (2022) who inserts a single sequence of prompt tokens at different positions, we insert at least one and at most n prompt series per concatenation, yielding a total of $m \cdot (2^n - 1)$ prompt positions. For the **Prefix style** prompts, we explore n insertion points ($n = 2$ and $n = 3$ corresponding to the first row in subfigure 1a and subfigure 1b) without the [mask] token ($m = 1$), which results in $2^n - 1$ different prompt positions.

[2]Regarding the sentence-pair classification task, we maintain the expected task input sequence order to narrow our focus on the prompt position (e.g. premise then hypothesis).



(a) single-sentence tasks

(b) sentence-pair tasks

Figure 1: Insertion positions for cloze-style prompts

4 Effect on Gradient-based Prompting

4.1 Setup

For gradient-based approaches, both discrete [3] and continuous methods are investigated. To focus on studying the effect of prompt position themselves, we implement two vanilla approaches:

Prompt-based fine-tuning: For discrete prompt, we fine-tune all the LM's parameters with the input restructured within a manual prompt template as per (Schick and Schütze, 2020; Gao et al., 2020).

Prompt tuning: For continuous prompt, we instantiate standard prompt tuning (Lester et al., 2021), which only tunes the continuous prompts tokens prepended to the input layer with the language model frozen. Besides, we incorporate both cloze and prefix styles, leading to four types of prompts for empirical investigation: ***cloze manual prompt***, ***cloze continuous prompt***, ***prefix manual prompt*** and ***prefix continuous prompt***.

Models: We choose language models which are popular in the NLU research literature. As per Gao et al. (2020), we use Roberta-large (Liu et al., 2019) to predict the masked token based on the cloze-style prompt. To generate answers from prefix-style prompts, we use T5-large language model adaption as per Lester et al. (2021), which is pre-trained for 10K steps with language modelling objectives without mixing downstream tasks. We additionally experiment with T5-XL (3B) on prefix continuous prompt to investigate the relatively larger model. All our few-shot experiments are conducted using the OpenPrompt framework[4](Ding et al., 2021).

Datasets: We examine the above approaches on five commonly used natural language understanding datasets as per (Gao et al., 2020; Lester et al., 2021). The datasets span various tasks: sentiment analysis (CR (Hu and Liu, 2004), SST-2 (Wang et al., 2018)), question classification (TREC (Voorhees and Tice, 2000)), question-answering (Boolq (Wang et al., 2019)), and natural language inference (RTE (Wang et al., 2019)), broadly classified into single-sentence (SST-2, CR, TREC) and sentence-pair categories (RTE, Boolq). For the datasets from GLUE and SuperGLUE, we use the original development sets for testing, and for the rest, we follow the testing sets as per Gao et al. (2020). See Table 9 in Appendix for details.

We measure the effect of the prompt position by the model's few-shot performance. We construct D_{train} and D_{dev} with K samples per label from the original training data, with K ranging from 16 to 128. We calculate the average accuracy across five randomly sampled D_{train} and D_{dev} splits, using the same fixed set of seeds as per Gao et al. (2020).

Positions: As explained in Section 3.2, we experiment with distinct prompt positions tailored to different prompt-style: 14 positions for cloze-style prompts in single-sentence tasks and 45 in sentence-pair tasks; for prefix-style prompts, 3 positions in single-sentence tasks and 7 in sentence-pair tasks.

To mitigate the influence of vocabulary in the manual prompt, we employ a reference template from prior research and only change the position where these words are inserted (See Table 2 for simple examples). In practice, there would be different situations for each prompt position, particularly with discrete prompts. For example, when inserting task input between two prompt sequences, the choice of which sequence to place at the beginning or end can vary. We prioritize templates that maintain grammatical coherence, selecting one template per prompt position[5]. With regards to continuous prompts, when inserting multi-prompt series, we simply separate continuous tokens used in the single prompt sequence equally to mitigate the effect of prompt length. All templates and verbalizers

[3] We do not take the automated search approach in discrete space, as this may result in different vocabulary in templates and potentially obfuscate the impact prompt position has in our results.

[4] https://github.com/thunlp/OpenPrompt

[5] All prompts obtained by altering position presented in this paper were chosen before evaluation to avoid any selection bias.

		SST-2		CR		TREC		RTE		Boolq	
K-size	Method	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)
16	CM	5.1	90.8	3.3	91.2	3.3	85.7 (+2.9)	17.8	71.2 (+3.7)	14.8	69.5 (+2.0)
	CC	23.1	86.3 (+0.2)	18.7	84.6 (+0.8)	17.4	65.1 (+6.8)	10.2	57.0 (+1.4)	17.1	61.1 (+1.0)
	PM	3.8	89.5	2.7	92.7 (+0.2)	0.8	85.0 (+0.3)	6.2	61.7 (+3.1)	6.7	64.2 (+3.0)
	PC	13.7	70.6 (+13.4)	26.4	86.8	3.9	71.8 (+2.4)	4.4	52.1	4.3	49.6 (+4.3)
	PC-XL	12	84.9	18.3	86.7	5.5	80.8 (+5.5)	6.6	53.1 (+1.7)	12.1	51.8 (+4.8)
32	CM	4.4	91.4 (+0.7)	2.5	92.0 (+0.9)	1.6	88.8 (+1.3)	20.6	74.7 (+1.8)	17	72.1 (+3.8)
	CC	17.9	89.7	14.8	89.2 (+0.2)	12.9	74.8 (+1.7)	10.5	59.6 (+0.5)	14.9	63.5 (+6.1)
	PM	0.9	92	1.9	93.0 (+0.1)	0.4	88.3 (+0.4)	11.4	66.9 (+7.0)	6.3	71.1
	PC	12.8	75.5 (+8.4)	17	89.4	5.6	80.4 (+5.6)	2.5	52.6 (+0.2)	7.9	57.3 (+5.2)
	PC-XL	4.1	90.6	2.1	90.4 (+0.4)	2.8	84.9 (+2.8)	5.3	54.4	5.3	53.7 (+4.8)
64	CM	2.4	92.4 (+0.9)	1.6	92.3 (+0.2)	1.3	92.3 (+1.1)	22.9	77.3 (+4.3)	14	73.8 (+0.6)
	CC	10.9	91.4 (+1.2)	7.8	91.1 (+1.3)	10.1	81.8	14.4	63.5 (+4.3)	11.4	62.6 (+0.5)
	PM	0.7	93.1 (+0.7)	0.6	93.7	0.4	90.5	12.9	72.9 (+4.8)	3.1	73.8
	PC	14	89.1 (+5.3)	13.3	92	3.5	85.5 (+3.5)	6.8	57.3	4.9	55
	PC-XL	1.3	92.8	0.8	92.7 (+0.8)	1.4	88.6	10.4	62.3	6.1	55.1 (+2.5)
128	CM	1.7	93	1.2	92.6	1.1	94.6 (+0.7)	18.1	79.8 (+1.0)	13.3	75.7
	CC	7.6	92.6 (+2.3)	4.8	92.1 (+1.4)	6.2	88.1 (+0.7)	12.2	63.1 (+3.0)	13.9	65.1 (+4.7)
	PM	0.6	93.4 (+0.2)	0.4	93.9 (+0.2)	0.6	92.2 (+0.6)	9	75.2 (+3.8)	2.7	76.5 (+1.0)
	PC	7.7	88.5	5.7	91.6	1.6	89.4 (+1.6)	9.6	61.2 (+2.1)	5.9	54.8 (+3.4)
	PC-XL	1.4	93.3	0.5	93.9	1	91.7 (+0.5)	9.9	64.4	10.5	58.2 (+5.4)

Table 1: Impact of Prompt Position in the few-shot gradient-based prompting across K training samples per label. 'Var' indicates the performance disparity between the least and most effective positions. 'Best' is the accuracy at the optimal position. (Δ) is the accuracy delta of the optimal position compared to the reference position. If the optimal position is the reference position, then (Δ) is not reported. We use the following abbreviations. **CM**: Cloze manual prompt; **CC**: Cloze continuous prompt; **PM**: Prefix manual prompt; **PC**: Prefix continuous prompt; **PC-XL**: Prefix continuous prompt with the T5-XL model.

Prefix manual prompt
Question: Is this sentence positive or negative? Answer: {text}
{text} Question: Is this sentence positive or negative? Answer:
Question: Is this sentence positive or negative? {text} Answer:

Table 2: Example templates with different positions for SST-2. The italic row indicates the default template position in Gao et al. (2021).

we used are described in Appendix C, along with their respective prompt position options for different tasks (Appendix D).

4.2 Results

As demonstrated in Table 1, for single-sentence tasks - SST-2, CR and TREC, the influence of prompt position is relatively small when using manual prompts, while significant performance variations arise when continuous prompts are employed. With the K-size increases, the differences between all methods tend to diminish. We note that the optimal prompt position may not always align with the reference position across different methods especially in the TREC dataset. Yet, the accuracy gap compared to the reference position also becomes smaller when K is set to 128.

For sentence-pair tasks, RTE and Boolq, a substantial performance variation is observed across all methods. As the K increases, the variance between different prompt positions persists except for the case of the prefix manual prompt in Boolq. Similar to single-sentence tasks, the reference prompt position does not consistently produce optimal results across all methods. Notably, even when the K is set to 128, there are instances where a noticeable difference exists between the best-performing prompt position and the reference position. For example, in RTE, the prefix manual prompt shows a 3.8 percentage point difference in performance, and in Boolq, the prefix continuous prompt has a 5.4 percentage point difference (Table 1).

We additionally experiment with T5-3B on prefix continuous prompt, more commonly used in prompt tuning methods, to investigate the effects of prompt positions in a relatively large model. Our results indicate that a larger scale helps to reduce the performance difference in single-sentence tasks, especially when $K > 16$. However, the variance remains considerable in sentence-pair tasks.

In general, sentence-pair tasks are more susceptible to the influence of prompt position compared to single-sentence tasks, whereas continuous prompts exhibit higher sensitivity to position compared to manual prompts.

K-Shot	Method	RTE	Boolq
K=16	CM	the Answer: {text_a} . [mask] {text_b} Question: ?	Answer: {text_a} . Question: {text_b} ? [mask] .
	CC	***P*** {text_a} ***P*** [mask] {text_b} ***P***	***P*** {text_a} ***P*** [mask] {text_b}
	PM	{text_a} Question: True or False? Answer: {text_b}	Question: {text_a} {text_b} ? Answer:
	PC	***P*** {text_a} {text_b}	***P*** {text_a} ***P*** {text_b}
	PC-XL	{text_a} ***P*** {text_b} ***P***	{text_a} ***P*** {text_b}
K=32	CM	the Answer: {text_a} . [mask] {text_b} Question: ?	{text_a} . Question: ? Answer: . [mask] {text_b}
	CC	***P*** {text_a} [mask] {text_b} ***P***	{text_a} [mask] {text_b} ***P***
	PM	Question: True or False? Answer: {text_a} {text_b}	{text_a} Question: {text_b} ? Answer:
	PC	***P*** {text_a} ***P*** {text_b}	***P*** {text_a} ***P*** {text_b}
	PC-XL	***P*** {text_a} {text_b}	{text_a} ***P*** {text_b}
K=64	CM	the Answer: {text_a} . [mask] {text_b} Question: ?	{text_a} . Question: ? {text_b} [mask] Answer: .
	CC	{text_a} [mask] {text_b} ***P***	{text_a} [mask] {text_b} ***P***
	PM	Answer: {text_a} Question: True or False? {text_b}	{text_a} Question: {text_b} ? Answer:
	PC	***P*** {text_a} {text_b}	{text_a} {text_b} ***P***
	PC-XL	***P*** {text_a} {text_b}	***P*** {text_a} ***P*** {text_b}
K=128	CM	the Answer: . {text_a} Question: ? {text_b} [mask]	{text_a} . Question: {text_b} ? Answer: [mask] .
	CC	***P*** {text_a} [mask] {text_b} ***P***	{text_a} [mask] {text_b} ***P***
	PM	True or False? {text_a} Question: {text_b} Answer:	{text_a} Question: ? Answer: {text_b}
	PC	{text_a} {text_b} ***P***	***P*** {text_a} ***P*** {text_b} ***P***
	PC-XL	***P*** {text_a} {text_b}	***P*** {text_a} ***P*** {text_b} ***P***

Table 3: Optimal prompt position on RTE and Boolq. ***P*** denotes the soft prompt tokens.

4.3 Discussion

In our main paper, we detail the optimal prompt positions for sentence-pair tasks in Table 3, and for single-sentence tasks in Table 11, found in the Appendix. Table 3 shows that the optimal prompt position is not consistently shared across different datasets when employing the same prompt method. For example, for the prefix continuous prompting (PC), the RTE task often prefers prompt tokens inserted at the start, especially with the T5-3B model. In contrast, Boolq shows a preference for the position "***P*** {text_a} ***P*** {text_b} ***P***" in both T5-large and T5-XL models (K=128). Also, we have noticed that the optimal position varies depending on K size, indicating that the distribution of input samples holds an influence.

There is no clear superiority between inserting multiple prompt sequences and a single prompt sequence. For example, RTE appears to respond better to a single-sequence prompt in prefix continuous prompting, while Boolq tends to favour multiple sequences. However, the relative position of the [mask] token in cloze-style prompts indeed affects the model performance, which is consistent with the findings of Gao et al. (2020). Typically, placing the [mask] token between the two inputs is often favoured in RTE. We conduct supplementary experiments with a null template which will be further elaborated in Appendix A.

It is worth noting that grammar doesn't always dictate the performance of manual prompts. This can be observed where grammatically incorrect prompts often achieve the best performance, and the performance difference between grammatically correct and incorrect prompts is not always negligible. For example, in the Boolq dataset within a cloze-style manual prompt (K = 32), *"{text_a} . Question: ? Answer: . [mask] {text_b}"* (shown in Table 3) outperforms the reference prompt *"{text_a} . Question: {text_b} ? Answer: [mask] ."* (Schick and Schütze, 2020) by 3.84 percentage points (Table 1). This phenomenon suggests that prompts considered reasonable by humans may not necessarily be effective for language models, which is consistent with Liu et al. (2021c). It implies that factors beyond grammar contribute to performance outcomes.

5 Effect on Gradient-free Prompting

5.1 Setup

We explore classic gradient-free prompting methods within both zero-shot and few-shot paradigms. For few-shot settings, in-context learning is investigated via **direct prompting** (Brown et al., 2020) as well as **chain-of-thought** (CoT) prompting, where models provide a reasoning step prior to the final response (Wei et al., 2022b). For zero-shot settings, we only consider the direct prompting.

We examine the in-context learning performance of SST-2 using a 16-shot setting, which is the maximum length accommodated within 512 tokens, on the T5-large model. We find a consistent trend that the position variance is significant and the reference position is sub-optimal. However, the overall performance of in-context learning significantly lags behind that of the fine-tuning method by em-

	T5-3B		Flan-T5-3B		T5-11B		Flan-T5-11B		LLaMA-13B		Flan-LLaMA-13B	
	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)
Causal Judgement	4.3	48.7	0.5	63.1	1.6	50.3 (+1.6)	0.5	60.4 (+0.5)	1.6	48.1	3.2	59.9 (+3.2)
Disambiguation QA	1.6	33.2	1.2	68.8 (+1.2)	5.2	34.8	3.2	69.2 (+2.8)	24.0	60.0	10	64.8
Sports Understanding	2.0	48.0	0.8	56.0 (+0.8)	2.8	51.6 (+2.8)	2.8	70.4	2.4	62.8 (+2.4)	2.4	67.6
Navigate	14	42	4.8	61.2	7.2	47.6 (+2.8)	2.4	61.2 (+2.4)	0	58	7.6	44.4 (+2.8)
Logical Deduction (5)	0	22.4	6.4	48.4 (+1.6)	4.8	23.2 (+3.2)	6.4	57.2 (+1.6)	6.4	26.8 (+3.6)	10.4	37.2
Logical Deduction (7)	1.6	15.2 (+1.6)	6.8	52.8	1.2	18.8 (+0.4)	5.6	60.8 (+1.2)	0.8	15.6 (+0.8)	10.4	32.0 (+1.2)
Logical Deduction (3)	36.4	36.4	4.8	62.4 (+0.8)	1.6	36	0.4	72	6.8	40.8	1.6	45.2 (+1.6)
Penguins in a Table	8.2	24.0 (+8.2)	5.5	37.7	4.1	25.3 (+3.4)	9.6	44.5 (+2.0)	0	26.7	10.3	37
Salient Translation Err.	4	16.0 (+4)	11.2	45.2	0	12	6.8	51.2	12.8	35.6 (+0.4)	14.8	36.8 (+1.2)
Movie Recommendation	2.0	27.2 (+2.0)	20.8	65.2 (+10)	15.6	44	32.8	71.6 (+11.6)	31.6	77.2 (+4)	19.2	81.6 (+12.8)
All task (avg)	7.41		6.28		4.41		7.05		8.64		8.99	

Table 4: The effect of prompt position in few-shot direct prompting on BBH benchmark.

	T5-3B		Flan-T5-3B		T5-11B		Flan-T5-11B		LLaMA-13B		Flan-LLaMA-13B	
	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)
Causal Judgement	0.5	43.3	5.9	60.4 (+3.2)	2.6	48.1	0.5	57.2 (+0.5)	2.7	54	5.4	57.8 (+2.7)
Disambiguation QA	2.4	28.4	1.6	69.2	2	30.0 (+0.8)	1.2	64.8 (+0.8)	8.4	50.0 (+8.4)	1.6	39.2 (+1.6)
Sports Understanding	1.2	46.8 (+0.8)	5.2	58.0 (+2.0)	4	59.2 (+4.0)	4.4	64.0 (+2.8)	1.6	78.4 (+1.6)	2.8	77.2
Navigate	5.2	27.6 (+0.8)	6.8	56.8 (+0.4)	4.4	32.8	4	61.6	2.8	59.6 (+0.8)	2.4	57.6
Logical Deduction (5)	4.0	20.8 (+4.0)	10	37.6	8.4	21.6 (+1.2)	9.2	51.2	1.2	29.6	4.4	32.8 (+4.4)
Logical Deduction (7)	3.2	14.4 (+1.6)	12	30	6.4	17.2	18.8	53.6	3.2	22.8 (+1.2)	0.8	20.0 (+0.4)
Logical Deduction (3)	2.8	38.8 (+2.8)	2.8	55.6	8.4	40.4	11.2	67.2	3.6	49.2	2.4	57.6 (+2.0)
Penguins in a Table	4.2	29.5 (+4.2)	2.7	26.7 (+1.4)	6.9	24	13.7	43.8 (+6.8)	5.5	40.4	3.5	42.5
Salient Translation Err.	2.4	9.2	7.6	25.6 (+2.0)	8	14	27.2	38.4	5.2	20.4 (+5.2)	0	30.8
Movie Recommendation	3.6	26	15.6	53.2	18.4	47.2	11.2	44.4	15.2	70 (+15.2)	25.6	74
All task (avg)	2.95		7.02		6.95		10.14		4.94		4.89	

Table 5: The effect of prompt position in few-shot COT prompting on BBH benchmark.

ploying the same manual prompt (see Table 18 in Appendix). Consequently, we shift our focus to relatively larger models in this section.

Models: We investigate two T5-series models, T5-3B and T5-11B (Raffel et al., 2023), and a decoder-only model LLaMA-13B (Touvron et al., 2023), which have sufficient capability to assess the impact of prompt position without any fine-tuning. Additionally, we conduct experiments with their instruction-tuned variants, Flan-T5-XXL (Chung et al., 2022) and Flan-LLaMA (Wang et al., 2023). These variants have been pre-trained on a diverse set of data sources utilizing an array of instruction template types that incorporate a wide spectrum of vocabulary and positional variations.

For evaluation, we measure accuracy using the exact match in the few-shot direct prompting as per Suzgun et al. (2022); Chung et al. (2022). Within the CoT setup, we extract the first word after the phrase *'[T|t]he answer is'*, or capture the full response if no such pattern is present. For the zero-shot setting, we compute the likelihood of each target option and select the option with the highest log-likelihood as per Brown et al. (2020); Sanh et al. (2022); Wei et al. (2022a).

Datasets: We evaluate the sub-tasks of BIG-Bench Hard (BBH) (Suzgun et al., 2022), a challenging benchmark from BIG-Bench[6], for the fact that instruction-tuned models were not exposed to it during training. The tasks involve not only NLU (e.g. Disambiguation QA) but also reasoning (e.g. Navigate, Logical Deduction), the Use of World Knowledge (e.g. Sports Understanding, Movie Recommendation, Causal Judgement), Table Parsing (e.g. Penguins in a Table) and the Multilingual task (e.g. Salient Translation Error detection). Following Suzgun et al. (2022), we employ the officially provided prompts, each accompanied by three few-shot examplers, for both Chain-of-Thought and Direct prompting.

Positions: This experiment is constrained to one prompt type, the prefix manual prompt, a choice informed by the nature of the models and methods we employed here. Given the only one task-input structure of the BBH benchmark, we investigate

[6]We choose the tasks with defined task specifications from bench authors (2023) to ensure that we can effectively alter prompt positions.

	Flan-T5-3B		Flan-T5-11B		LLaMA-13B		Flan-LLaMA-13B	
	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)	Var	Best (Δ)
Causal Judgement	1.1	61	1.6	61.5 (+0.5)	2.7	53.5 (+1.6)	5.3	58.8 (+3.2)
Disambiguation QA	2.4	69.2 (+2.4)	4	68.8	20.8	52.8	16.4	60
Sports Understanding	2.4	53.6 (+1.6)	2.4	61.6 (+2.4)	8.8	60.8	5.6	64.4
Navigate	3.2	60.8 (+2.4)	1.6	60.8	7.6	49.6 (+7.6)	1.2	59.2 (+1.2)
Logical Deduction (5)	4	50.4 (+0.4)	7.2	56.8 (+1.6)	3.2	21.2	15.2	38
Logical Deduction (7)	6	53.2 (+0.8)	8	62.4 (+1.2)	0	14.8	15.2	33.6
Logical Deduction (3)	4.4	64.8	5.6	75.2	7.6	39.6 (+3.2)	7.2	50 (+0.4)
Penguins in a Table	8.9	41.1	9.5	47.9 (+6.1)	10.2	26	8.9	31.5
Salient Translation Err	3.2	42.4	0	50	4.4	18 (+4.0)	5.2	24.8
Movie Recommendation	28.8	67.2 (+13.2)	14.8	55.2 (+8.8)	34.4	56	38	67.6 (+2.0)
All task (avg)	6.44		5.47		9.97		11.82	

Table 6: The effect of prompt position in zero-shot setting on BBH benchmark

three prompt positions for each sub-task: insertion at the front, the rear, or on both sides of the input. We play the relative position between input and prompt within the exemplar delimiters (e.g., "Q:"/"A:")[7]. All templates and their variants of positions are detailed in Appendix C.

5.2 Results

Few-shot: As illustrated in Table 4 and Table 5, both direct prompting and COT prompting exhibit significant performance variances across different prompt positions in most scenarios. Also, the optimal position does not always align with the default settings. Notably, the performance disparity between the default (used in Suzgun et al. (2022)) and the best outcome (obtained by altering the position) is significant sometimes. For instance, in the movie recommendation, there is a 12.8 per cent performance gap for Flan-LLaMA under direct prompting, and a 15.2 per cent gap for LLaMA-13B in the CoT approach (changing from "*Find a movie similar to {Input}* \n*{Options}*" to "*{Options}* \n *Find a movie similar to {Input}*").

When comparing larger model scales (3B VS 11B), the variance persists, and in some cases, it even increases. For instance, in the task of three-object logical deduction, the variance among different positions rises from 2.8 per cent to 11.2 per cent on Flan T5-based models using the chain-of-thought method (Table 5). Similarly, and quite surprisingly, we observe that instruction tuning does not necessarily reduce the performance difference between positions; in fact, it can sometimes exacerbate them.

[7]Our rationale for this choice is that templates with exemplar delimiters consistently yield better results and also align with the approach of Chung et al. (2022) in the creation of few-shot templates for instruction fine-tuning.

Zero-shot: In zero-shot scenarios, T5 plain models show only marginal improvements or even worse than random, so our attention is primarily on the other four models. As indicated in Table 6, we observe similar trends to those in the few-shot setting. There are noticeable performance differences between various prompt positions in most cases, and the position yielding the best performance does not consistently match the reference position.

5.3 Discussion

In Table 7, we present the optimal prompt positions for various models across different tasks. It reveals that there is no universal best position across tasks, models and methods, echoing our findings discussed in Section 4.3. While the default prompt positions are typically set to "both" (namely multiple prompt sequences are inserted both in the front and rear), this doesn't yield the best performance all the time. Besides, the optimal position in the zero-shot setting may not correspond to the one in the few-shot setting. This is nuanced and consistent with our discussion in Section 4.3 that the input distribution may hold influence. Interestingly, within the same dataset domain (e.g., three logical deduction tasks), we observe a higher overlap of optimal positions within the same method and model. This is particularly notable in Flan-T5-series models, where "both" is the preferred position in the chain-of-thought approach across all three sub-tasks.

The grammar tends to better-preserved compared to the cases setup in Section 4 when we alter positions here. Despite this, there is still a significant variance in performance between different positions. We suspect this variance may stem from the "favourite" position of the language model - similarly formatted data encountered during pre-training or instruction-tuning. As previously noted,

	T5-3B			Flan-T5-3B			T5-11B			Flan-T5-11B			LLaMA-13B			Flan-LLaMA-13B		
	Direct	COT	0-shot	Direct	COT	0-shot	Direct	COT	0-shot	Direct	COT	0-shot	Direct	COT	0-shot	Direct	COT	0-shot
Causal Judgement	B	B	R	B	R	B/R	F	B	F/B	F/R	R	R	B	B	F	R	F	F
Disambiguation QA	B	B	F/R	F	B	F	B	R	F	F	F	B	B	F	B	B	F/R	B
Sports Understanding	F	B	-	B	R	B	R	R	R	F	R	R	R	B	F	F	F	F
Navigate	F/B	F	-	B	F	R	F	B	-	R	B	B	-	F	R	R	B	R
Logical Deduction (5)	-	F	R	R	B	R	F	R	-	R	B	R	R	B/R	B	B	R	B
Logical Deduction (7)	F	F	R	B	B	R	R	B	B	R	B	R	R	F	-	R	F	B
Logical Deduction (3)	B	F	B	R	B	B/R	B	B	-	F/B	B	B	B	B	R	F	R	R
Penguins in a Table	F	R	F	B	F	B	F	F/B	F	F	F	F	-	B	B	B	B	B
Salient Translation Err.	R	B	F/B	B	R	B	-	B	R	F/B	B	-	R	F	F	F	-	B
Movie Recommendation	F	B	B	R	B	R	B	B	F	R	B	R	R	F	B	F	B	R

Table 7: The optimal position of different models on BBH. 'F' for Front Insertion, 'B' for Both Sides and 'R' for Rear Placement.

instruction-tuned models do not necessarily diminish this variance, despite typically incorporating a wide range of vocabulary and positional variations in their training templates. We have looked at their ten templates for each dataset in Chung et al. (2022) and found an uneven distribution of position variety. For instance, in the SST-2 tasks, a majority of their instruction tuning templates position input sentences between the task specification and options. In contrast, in the CNN Daily Mail summarization task (See et al., 2017), the templates commonly place input texts at the start.

To further explore the potential reason for this variance, we conduct a preliminary experiment on SST-2 using Flan-based models under the zero-shot setting (Table 19 in Appendix). We find that all models exhibit strong performance with minimal variance or even zero (e.g. Flan-T5-11B) between different positions. This outcome is expected due to the exposure of SST-2's training data during the instruction-tuning process, but it is also encouraging as it suggests that a strong language model should be robust to variations in prompt positioning. Nonetheless, we still observe a slight preference for the 'both' position in both Flan-T5-3B and Flan-LLaMA-13B models, supporting our hypothesis that one task may favour a prompt position more commonly encountered during the training process.

6 General Discussion

Our main research question in this paper is whether and how prompt positions matter. In Section 4, we observe that continuous prompts are more sensitive to prompt positions compared to manual prompts. In Section 5, all language models exhibit a certain degree of sensitivity to positional changing in both few-shot and zero-shot settings. Both sections 4 and 5 highlight that the optimal prompt position is not shared across tasks, and sometimes even differs among items of data. Furthermore, Section 5 reveals that instruction-tuning fails to mitigate positional variance, possibly due to an unequal distribution of position variety in training templates.

Our goal is not to identify a 'best' position to replace the prompt position used in prior works. Rather, we aim to highlight the effect of prompt position, which is often overlooked before. For gradient-based prompting such as prompt tuning, due to the influence of training data sizes, we suggest instance-dependent prompt position optimization as a valuable direction, with the potential to enhance model performance, which is similar to Wu et al. (2022); Yang et al. (2023). As for gradient-free prompting, in real-world scenarios, prompt positioning and the logic in which questions or instructions are structured is diverse. A robust language model should display a consistent understanding of inputs with the same semantic essence when only the prompt position changes. A future work direction is to explore if increasing the position variety during the instruction tuning process improves robustness of trained models.

7 Conclusion

In this study, we evaluated various prompt positions using both gradient-based and gradient-free approaches. Our findings reveal significant variations in performance based on prompt position across different methods, tasks, and models. Additionally, our research indicates that prompt positions commonly adopted in the existing literature often result in sub-optimal performance, with no single prompt position universally excelling across all tasks. These findings suggest prompt position optimization as a promising new direction in prompt engineering and advocate for the consideration of position-aware instruction tuning to develop more robust models in the future.

8 Limitations

Due to the extensive workload of experiments, we only test our hypothesis for 5 sub-tasks from (Gao et al. (2020); Lester et al. (2021)) in gradient-based approaches and 10 sub-tasks from Suzgun et al. (2022) in gradient-free approaches. We use medium-sized language models in all our experiments (e.g. LLaMA-13B) using relatively low computational resources (i.e. a single GPU card for inferring or training), so although we strongly suspect the results will be the same in the latest Large Language Models (e.g. GPT4) this needs to be confirmed empirically in future experiments.

References

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. PADA: A prompt-based autoregressive approach for adaptation to unseen domains. *CoRR*, abs/2102.12206.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

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. *CoRR*, abs/2005.14165.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

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

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. Ppt: Pre-trained prompt tuning for few-shot learning.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Abdullatif Köksal, Timo Schick, and Hinrich Schütze. 2023. Meal: Stable and active learning for few-shot prompting.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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

Xiangyang Liu, Tianxiang Sun, Xuanjing Huang, and Xipeng Qiu. 2022. Late prompt tuning: A late prompt could be better than many prompts.

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

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021c. GPT understands, too. *CoRR*, abs/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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, pages 1–11.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

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

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

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*.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them.

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

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

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 preprint arXiv:1804.07461*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, VG Vydiswaran, and Hao Ma. 2022. Idpg: An instance-dependent prompt generation method. *arXiv preprint arXiv:2204.04497*.

Xianjun Yang, Wei Cheng, Xujiang Zhao, Linda Petzold, and Haifeng Chen. 2023. Dynamic prompting: A unified framework for prompt tuning.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Effect of the position [mask] token

We have discovered that the position of the [mask] token has an impact on the cloze-style prompt, namely within Masked Language Models. To investigate this further, we conduct null template experiments with a K size of 16, where we simply concatenate the inputs and the [mask] token without a prompt. By analyzing the results of the null template as presented in Table 8, we observe that in single-sentence tasks, placing text before the [mask] token generally leads to better performance. For sentence-pair tasks, placing [mask] before the text is relatively sub-optimal. Notably, for the RTE task, positioning [mask] token in the middle of the two original inputs proves to be more advantageous. This observation also aligns with the overall performance trend demonstrated in the complete set of results, which can be found in Appendix D.

Interestingly, even when the [mask] token is placed in the same relative position to the task inputs, the performance exhibits a noticeable difference depending on how to insert the prompt sequences. For instance, in the cloze continuous prompt experiment on RTE dataset with a K size of 128, when the [mask] token is placed in the middle of two task inputs, "*{text_a} [mask] {text_b}* ***P***" achieves a performance that is 6.28 higher compared to "*{text_a}* ***P*** *[mask] {text_b}*" (detailed in Table 31).

Dataset	Null template	Mean (std)
SST-2	{text_a} [mask]	89.40 (1.44)
	[mask] {text_a}	82.80 (5.56)
CR	{text_a} [mask]	90.66 (1.19)
	[mask] {text_a}	89.01 (1.76)
TREC	{text_a} [mask]	85.99 (1.82)
	[mask] {text_a}	83.58 (1.58)
RTE	{text_a} {text_b} [mask]	55.02 (7.16)
	{text_a} [mask] {text_b}	65.49 (3.58)
	[mask] {text_a} {text_b}	53.86 (5.58)
Boolq	{text_a} {text_b} [mask]	64.70 (3.09)
	{text_a} [mask] {text_b}	64.05 (4.13)
	[mask] {text_a} {text_b}	59.05 (1.57)

Table 8: Null template results for all datasets on K size of 16 for cloze manual prompt.

B Experimental Details

For prompt-based fine-tuning, we employ an AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2e-5 and a batch size of 8 for 1000 steps, validating the performance every 100 steps. For prompt tuning on the Roberta model, we follow the setting in Sun et al. (2022), using AdamW with a learning rate of 5e-4 and a batch size of 16 for 1000 epochs, with model performance validation every 100 steps. For prompt tuning on the T5 model, we adopt Adafactor (Shazeer and Stern, 2018) with a learning rate of 0.3 and a batch size of 16 for 1000 steps, evaluating the performance every 8 steps. The prompt length for all experiments is set to 50, initialized from the first 50 tokens embeddings of the pre-trained language model following Ding et al. (2021) as initializing from the language model's vocabulary often gives better results. To mitigate overfitting, we employ the strategy of early stopping across all experiments.

All our models were trained on a single RTX8000 with 48GB of memory. In the main training experiments, continuous prompt tuning took approximately 547 hours for T5-3B, 264 hours for T5-large, and 2269 hours for Roberta-large. In prompt-based fine-tuning, T5-large and Roberta-large took around 150 and 546 hours, respectively. For the gradient-free main experiments, zero-shot experiments took roughly 16 hours, few-shot direct prompting took about 70 hours, and chain-of-thought prompting took around 345 hours.

C Dataset, reference prompt and positions

For gradient-based approaches, the dataset statistics used are shown in Table 9, the reference prompt positions are shown in Table 10, and the best-performing prompt positions for single-tasks are shown in Table 11. The CR templates employed here are consistent with that of SST-2 for all methods, following the setting outlined in Gao et al. (2020). Regarding the prefix continuous prompt (PC) applied to the TREC dataset, we follow the prompt position setting provided by Lester et al. (2021), which is commonly used as the default prompt position for most continuous prompt methods. Besides, Gu et al. (2022) use "***P*** *{text_b} [mask] {text_a}*" (e.g. {text_a} is "passage" and {text_b} is "question" in Boolq task) as the prompt position in cloze continuous prompt method. To narrow our focus on the prompt position and ensure consistency with the expected task input sequence order in other methods (e.g. manual prompt), we modify their input orders in this specific case. For

gradient-free approaches, the reference prompt positions used in Suzgun et al. (2022) and their variants are shown in Table 12.

D All results

D.1 SST-2

Full SST-2 results are presented in Tables 13, 15, 14, 16, and 17. The detailed results of two preliminary experiments mentioned in Section 5 can be found in Tables 18 and 19.

D.2 CR

Full CR results can be found in Tables 20, 22, 21, 23, and 24.

D.3 TREC

Full TREC results can be found in Tables 25, 27, 26, 28 and 29.

D.4 RTE

Full RTE results can be found in Tables 30, 31, 32, 33 and 34.

D.5 Boolq

Full Boolq results can be found in Tables 35, 36, 37, 38 and 39.

D.6 BBH

For zero-shot, few-shot direct prompting, and few-shot chain-of-thought prompting, full results are shown in Table 42, Table 40, and Table 41.

Corpus	$\|Y\|$	Train	Validation	Task	Evaluation Metrics
Single Sentence Tasks					
CR	2	1775	2000	sentiment	accuracy
SST-2	2	67349	1821	sentiment	accuracy
TREC	6	5452	500	question cls.	accuracy
Sentence Pair Tasks					
RTE	2	2491	278	NLI	accuracy
boolq	2	9427	3270	QA	accuracy

Table 9: The datasets evaluated in our work. $|Y|$ represents the number of classes within each task. We only sample D_{train} and D_{dev} of $K \times |Y|$ examples from the original training set in the few-shot experiments.

Method	**Task**	**Template**	**Verbalizer**	**Reference**
CM	SST2	{text_a} It was [mask].	positive: great, negative: terrible	(Gao et al., 2020)
	CR	{text_a} It was [mask].	positive: great, negative: terrible	(Gao et al., 2020)
	TREC	{text_a} This question is related to [mask] category.	abbr.: Expression, entity: Entity, description: Description, human: Human, location: Location, numeric: Number	(Köksal et al., 2023)
	RTE	{text_a} Question: {text_b}? the Answer: [mask].	entailment: yes, not_entailment: no	(Liu et al., 2021c)
	Boolq	{text_a}. Question: {text_b}? Answer: [mask].	entailment: yes, not_entailment: no	(Schick and Schütze, 2020)
CC	SST2	prompt {text_a} [mask]	positive: great, negative: terrible	(Gu et al., 2022)
	CR	***P*** {text_a} [mask]	positive: great, negative: terrible	
	TREC	***P*** [mask] {text_a}	abbr.: Expression, entity: Entity, description: Description, human: Human, location: Location, numeric: Number	(Liu et al., 2022)
	RTE	***P*** {text_a} [mask] {text_b}	entailment: yes, not_entailment: no	(Gu et al., 2022)
	Boolq	***P*** {text_a} [mask] {text_b}	true: yes, false: no	(Gu et al., 2022)
PM	SST2	{text_a} Question: Is this sentence positive or negative? Answer:	positive: positive, negative: negative	(Gao et al., 2021)
	CR	{text_a} Question: Is this sentence positive or negative? Answer:	positive: positive, negative: negative	
	TREC	Categories: $\{',' .join(label_words)\}$. What category best describes: {text_a} Answer:	abbr.: Abbreviation, entity: Entity, description: Description, human: Person, location: Location, numeric: Quantity	(Sanh et al., 2022)
	RTE	{text_a} Question: {text_b} True or False? Answer:	entailment: true, not_entailment: false	(Brown et al., 2020)
	Boolq	{text_a} Question: {text_b}? Answer:	true: yes, false: no	(Brown et al., 2020)
PC	SST2	***P*** {text_a}	positive: positive, negative: negative	(Lester et al., 2021)
	CR	***P*** {text_a}	positive: positive, negative: negative	
	TREC	***P*** {text_a}	abbr.: Abbreviation, entity: Entity, description: Description, human: Person, location: Location, numeric: Quantity	
	RTE	***P*** {text_a} {text_b}	entailment: entailment, not_entailment: contradiction	(Lester et al., 2021)
	Boolq	***P*** {text_a} {text_b}	true: true, false: false	(Lester et al., 2021)

Table 10: All reference prompt positions used in the main text of the paper. ***P*** denotes the continuous prompt tokens.

K-Shot	Method	SST-2	CR	TREC
K=16	CM	{text_a} It was [mask].	[mask] It was {text_a}.	{text_a} [mask] This question is related to category
	CC	***P*** [mask] ***P*** {text_a}	***P*** {text_a} ***P*** [mask]	***P*** {text_a} [mask]
	PM	{text_a} *Q*: Is this sentence positive or negative? *A*:	*Q*: Is this sentence positive or negative? {text_a} *A*:	*A*: *C*: $\{',' .join(label_words)\}$. What category best describes: {text_a}
	PC	{text_a} ***P***	***P*** {text_a}	***P*** {text_a} ***P***
	PC-XL	***P*** {text_a}	***P*** {text_a}	{text_a} ***P***
K=32	CM	It was [mask]. {text_a}	{text_a}. It was [mask]	This question is related to category. {text_a} [mask]
	CC	***P*** {text_a} [mask]	[mask] ***P*** {text_a}	***P*** [mask] {text_a} ***P***
	PM	{text_a} *Q*: Is this sentence positive or negative? *A*:	*Q*: Is this sentence positive or negative? {text_a} *A*:	{text_a} *C*: $\{',' .join(label_words)\}$. What category best describes: *A*:
	PC	{text_a} ***P***	***P*** {text_a}	***P*** {text_a} ***P***
	PC-XL	***P*** {text_a}	***P*** {text_a} ***P***	{text_a} ***P***
K=64	CM	[mask] It was {text_a}.	It was. {text_a} [mask]	{text_a} [mask] This question is related to category
	CC	***P*** [mask] ***P*** {text_a}	{text_a} [mask] ***P***	***P*** [mask] {text_a}
	PM	*Q*: Is this sentence positive or negative? *A*: {text_a}	{text_a}*Q*: Is this sentence positive or negative? *A*:	*C*: $\{',' .join(label_words)\}$. What category best describes: {text_a} *A*:
	PC	{text_a} ***P***	***P*** {text_a}	{text_a} ***P***
	PC-XL	***P*** {text_a}	***P*** {text_a} ***P***	***P*** {text_a}
K=128	CM	It was [mask]. {text_a}	{text_a} It was [mask].	{text_a} [mask] This question is related to category
	CC	***P*** [mask] ***P*** {text_a}	***P*** {text_a} ***P*** [mask]	***P*** {text_a} [mask] ***P***
	PM	*Q*: Is this sentence positive or negative? *A*: {text_a}	*Q*: Is this sentence positive or negative? *A*: {text_a}	*A*: *C*: $\{',' .join(label_words)\}$. What category best describes: {text_a}
	PC	***P*** {text_a}	***P*** {text_a}	***P*** {text_a} ***P***
	PC-XL	***P*** {text_a}	***P*** {text_a}	***P*** {text_a} ***P***

Table 11: Best-performing prompt position for single-sentence tasks with SST-2, CR and TREC datasets. ***P*** denotes the continuous prompt tokens, while *Q*, *A*, and *C* represent the abbreviations for "Question", "Answer", and "Categories", respectively. $\{',' .join(label_words)\}$ simplifies the representation of six label words used in the TREC dataset.

Causal Judgement	F	Q: How would a typical person answer each of the following questions about causation?\n{Options}\n{Input}\nA:
	B*	Q: How would a typical person answer each of the following questions about causation?\n{Input}\n{Options}\nA:
	R	Q: {Input}\n{Options}\nHow would a typical person answer each of the following questions about causation?\nA:
Disambiguation QA	F	Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.\n{Options}\n{Input}\nA:
	B*	Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.\n{Input}\n{Options}\nA:
	R	Q: {Input}\nIn the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.\n{Options}\nA:
Sports Understanding	F*	Q: Is the following sentence plausible? {Input}\nA:
	B	Q: Is the following sentence {Input} plausible?\nA:
	R	Q: {Input} Is the following sentence plausible?\nA:
Navigate	F	Q: If you follow these instructions, do you return to the starting point?\n{Options}\n{Input}\nA:
	B*	Q: If you follow these instructions, do you return to the starting point? {Input}\n{Options}\nA:
	R	Q: {Input} If you follow these instructions, do you return to the starting point?\n{Options}\nA:
Logical Deduction (3,5,7)	F	Q: The following paragraphs each describe a set of three (five/seven) objects arranged in a fixed order. The statements are logically consistent within each paragraph.\n{Options}\n{Input}\nA:
	B*	Q: The following paragraphs each describe a set of three (five/seven) objects arranged in a fixed order. The statements are logically consistent within each paragraph. {Input}\n{Options}\nA:
	R	Q: {Input} The following paragraphs each describe a set of three (five/seven) objects arranged in a fixed order. The statements are logically consistent within each paragraph.\n{Options}\nA:
Penguins in a Table	F	Q: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm.\n{Options}\n{Input}\nA:
	B*	Q: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. {Input}\n{Options}\nA:
	R	Q: {Input} Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm.\n{Options}\nA:
Salient Translation Err.	F	Q: The following translations from German to English contain a particular error. That error will be one of the following types: Named Entities: An entity (names, places, locations, etc.) is changed to a different entity. Numerical Values: Numerical values (ordinals or cardinals), dates, and/or units are changed. Modifiers or Adjectives: The modifiers and adjectives pertaining to a noun are changed. Negation or Antonyms: Introduce or remove a negation or change comparatives to their antonyms. Facts: Trivial factual errors not pertaining to the above classes are introduced in the translations. Dropped Content: A significant clause in the translation is removed. Please identify that error.\n{Options}\n{Input}\nA:
	B*	Q: The following translations from German to English contain a particular error. That error will be one of the following types: Named Entities: An entity (names, places, locations, etc.) is changed to a different entity. Numerical Values: Numerical values (ordinals or cardinals), dates, and/or units are changed. Modifiers or Adjectives: The modifiers and adjectives pertaining to a noun are changed. Negation or Antonyms: Introduce or remove a negation or change comparatives to their antonyms. Facts: Trivial factual errors not pertaining to the above classes are introduced in the translations. Dropped Content: A significant clause in the translation is removed. Please identify that error. {Input}\n{Options}\nA:
	R	Q: {Input}\n{Options}\nThe following translations from German to English contain a particular error. That error will be one of the following types: Named Entities: An entity (names, places, locations, etc.) is changed to a different entity. Numerical Values: Numerical values (ordinals or cardinals), dates, and/or units are changed. Modifiers or Adjectives: The modifiers and adjectives pertaining to a noun are changed. Negation or Antonyms: Introduce or remove a negation or change comparatives to their antonyms. Facts: Trivial factual errors not pertaining to the above classes are introduced in the translations. Dropped Content: A significant clause in the translation is removed. Please identify that error.\nA:
Movie Recommendation	F	Q: {Options}\nFind a movie similar to {Input}\nA:
	B*	Q: Find a movie similar to {Input}\n{Options}\nA:
	R	Q: {Input} Find a movie similar to\n{Options}\nA:

Table 12: The position variants on BBH. * denote the reference position used in Suzgun et al. (2022). The distinction in few-shot settings and COT involves adding instructions and examples in the former, and additional rationales in the latter. Our prompts and inputs remain unchanged and adhere to bench authors (2023); Suzgun et al. (2022), with modifications only on their relative position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
It was. {text_a} [mask]	90.02(1.44)	91.15(1.95)	90.92(1.90)	91.81(0.99)
{text_a} [mask] It was.	88.62(1.69)	90.71(1.60)	91.31(1.69)	91.56(0.52)
{text_a}. It was [mask]	90.21(1.58)	89.63(1.84)	91.74(1.37)	92.87(0.93)
It was. [mask] {text_a}	85.67(5.36)	89.15(1.79)	89.95(2.16)	91.28(1.04)
[mask] {text_a} It was.	85.92(3.55)	87.04(1.35)	90.30(1.34)	91.35(1.36)
[mask]. It was {text_a}	89.84(1.42)	89.61(3.55)	91.54(1.29)	91.86(1.13)
variance	4.54	4.11	1.79	1.59
Two prompt sequences				
It was {text_a}. [mask]	87.78(2.91)	89.70(0.68)	91.19(1.28)	92.64(0.82)
It was {text_a} [mask].	89.86(1.28)	89.72(1.85)	91.38(1.36)	91.95(0.91)
{text_a} It was [mask].	**90.80(1.74)**	90.71(2.01)	91.49(0.73)	**92.96(0.72)**
It was [mask]. {text_a}	90.09(1.26)	**91.42(0.91)**	91.54(1.81)	**92.96(0.33)**
It was [mask] {text_a}.	87.75(1.39)	88.60(0.76)	91.86(0.48)	92.20(1.14)
[mask] It was {text_a}.	89.20(1.46)	90.53(1.58)	**92.39(0.64)**	92.27(1.03)
variance	3.05	2.82	1.2	1.01
Three prompt sequences				
It {text_a} was [mask].	88.97(2.97)	90.30(1.04)	91.93(1.01)	91.74(0.98)
It [mask] was {text_a}.	88.14(2.66)	88.67(2.05)	91.19(0.53)	92.73(0.54)
variance	0.83	1.63	0.74	0.99
variance all	5.13	4.38	2.44	1.68

Table 13: Cloze manual prompt on SST-2. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
Question: Is this sentence positive or negative? Answer: {text_a}	85.73(6.94)	91.93(0.45)	**93.05(0.60)**	**93.39(0.39)**
{text_a} Question: Is this sentence positive or negative? Answer:	**89.45(4.16)**	**91.97(0.78)**	92.41(0.94)	93.17(0.45)
variance	3.72	0.04	0.64	0.22
Two prompt sequences				
Question: Is this sentence positive or negative? {text_a} Answer:	85.94(10.00)	91.06(0.96)	92.41(0.94)	92.84(0.98)
variance all	3.72	0.91	0.64	0.55

Table 14: Prefix manual prompt on SST-2. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} [mask]*	86.12(3.28)	**89.72(0.59)**	90.18(0.80)	90.32(0.85)
{text_a} [mask] ***P***	82.64(4.62)	89.11(2.68)	90.50(0.68)	91.10(0.96)
{text_a} ***P*** [mask]	72.09(6.69)	76.93(9.09)	83.23(3.27)	89.17(0.91)
P [mask] {text_a}	80.99(1.39)	82.00(4.65)	85.00(1.72)	86.90(2.75)
[mask] {text_a} ***P***	72.02(5.58)	71.83(5.71)	80.50(2.63)	84.98(1.31)
[mask] ***P*** {text_a}	75.07(7.10)	85.99(2.85)	89.91(1.68)	90.46(2.14)
variance	14.1	17.89	10	6.12
Two prompt sequences				
P {text_a} ***P*** [mask]	79.93(4.31)	83.62(6.26)	89.01(1.64)	92.09(0.49)
P {text_a} [mask] ***P***	75.32(5.68)	82.06(9.97)	89.86(2.07)	91.38(1.50)
{text_a} ***P*** [mask] ***P***	69.75(8.36)	77.48(7.96)	84.91(3.15)	91.01(1.02)
P [mask] ***P*** {text_a}	**86.31(2.59)**	89.27(1.70)	**91.42(0.88)**	**92.57(0.38)**
P [mask] {text_a} ***P***	74.79(5.98)	79.36(2.03)	84.93(2.56)	88.44(1.12)
[mask] ***P*** {text_a} ***P***	63.17(7.58)	74.66(6.49)	86.24(3.59)	90.67(2.38)
variance	23.14	14.61	6.51	4.13
Three prompt sequences				
P {text_a} ***P*** [mask] ***P***	70.55(9.44)	76.12(8.93)	83.67(4.82)	89.22(3.18)
P [mask] ***P*** {text_a} ***P***	85.11(0.73)	85.73(6.58)	90.05(2.05)	91.95(0.82)
variance	14.56	9.61	6.38	2.73
variance all	23.14	17.89	10.92	7.59

Table 15: Cloze continuous prompt on SST-2. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	57.20(6.23)	67.11(15.21)	83.76(4.90)	**88.53(3.67)**
{text_a} ***P***	**70.64(16.85)**	**75.48(14.73)**	**89.11(2.46)**	86.81(8.46)
variance	13.44	8.37	5.35	1.72
Two prompt sequences				
P {text_a} ***P***	56.90(3.14)	62.71(7.22)	75.05(9.16)	80.80(3.13)
variance all	13.74	12.77	14.06	7.73

Table 16: Prefix continuous prompt with T5-Large on SST-2 Dataset. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	**84.93 (6.43)**	**90.57 (2.32)**	**92.82 (1.01)**	**93.33 (1.39)**
{text_a} P	78.85 (15.79)	86.51 (5.72)	91.51 (0.8)	91.93 (2.43)
variance	6.08	4.06	1.31	1.4
Two prompt sequences				
P {text_a} ***P***	72.87 (16.72)	88.33 (4.32)	92.32 (1.48)	92.11 (0.97)
variance all	12.06	4.06	1.31	1.4

Table 17: Prefix continuous prompt with T5-XL on SST-2. The italic row indicates the reference prompt position.

	PBF mean(std)	ICL mean(std)
Question: Is this sentence positive or negative?\nAnswer:\n{text_a} [mask]	82.94 (12.69)	61.79 (14.77)
Question: Is this sentence positive or negative?\n{text_a}\nAnswer: [mask]	90.34 (1.54)	55.99 (14.56)
{text_a}\nQuestion: Is this sentence positive or negative?\nAnswer: [mask]	89.36 (3.64)	50.28 (1.76)

Table 18: 16-shot preliminary experiments of T5-Large on SST-2. PBF denotes Prompt-based fine-tuning whereas ICL denotes In-context Learning. The italic row indicates the reference prompt position.

	Front	Both	Rear
Flan-T5-3B	95.30	**95.41**	95.30
Flan-T5-11B	95.53	95.53	95.53
Flan-LLaMA-13B	94.95	**95.07**	94.38

Table 19: The accuracy of different prompt positions with Flan-based models on SST-2 under the zero-shot setting.

	K=16 **mean(std)**	**K=32** **mean(std)**	**K=64** **mean(std)**	**K=128** **mean(std)**
One prompt sequence				
It was. {text_a} [mask]	89.86(1.63)	89.54(0.74)	**92.25(0.48)**	92.10(0.91)
{text_a} [mask] It was.	90.64(2.22)	90.76(0.35)	91.54(0.62)	92.19(0.83)
{text_a}. It was [mask]	90.93(1.74)	**91.98(0.69)**	91.90(0.86)	92.06(0.88)
It was. [mask] {text_a}	90.24(1.65)	91.25(1.65)	91.91(0.84)	91.36(0.82)
[mask] {text_a} It was.	89.29(4.11)	89.96(2.39)	91.13(1.43)	91.78(0.42)
[mask]. It was {text_a}	91.07(1.97)	91.41(0.96)	91.55(1.11)	92.05(0.39)
variance	1.78	2.44	1.12	0.83
Two prompt sequences				
It was {text_a}. [mask]	89.29(1.76)	90.59(1.07)	91.73(0.43)	91.85(1.08)
It was {text_a} [mask].	90.11(1.51)	91.33(0.80)	91.95(0.18)	91.76(0.74)
{text_a} It was [mask].	91.17(0.81)	91.13(0.79)	92.12(0.74)	**92.55(0.75)**
It was [mask]. {text_a}	91.17(0.61)	90.93(1.78)	92.03(0.44)	91.93(0.74)
It was [mask] {text_a}.	88.06(4.32)	90.20(1.00)	90.73(0.86)	91.55(0.40)
[mask] It was {text_a}.	**91.22(0.53)**	90.48(1.16)	91.32(0.47)	92.29(0.85)
variance	3.11	1.13	1.39	1
Three prompt sequences				
It {text_a} was [mask].	87.86(4.30)	90.31(0.74)	92.08(0.98)	91.94(0.49)
It [mask] was {text_a}.	88.86(1.23)	90.85(1.25)	91.54(0.72)	91.66(0.74)
variance	1	0.54	0.54	0.28
variance all:	3.36	2.44	1.52	1.19

Table 20: Cloze manual prompt on CR. The italic row indicates the reference prompt position.

	K=16 **mean(std)**	**K=32** **mean(std)**	**K=64** **mean(std)**	**K=128** **mean(std)**
One prompt sequence				
Question: Is this sentence positive or negative? Answer: {text_a}	89.98(0.38)	91.13(1.87)	93.06(1.03)	**93.91(0.38)**
{text_a} Question: Is this sentence positive or negative? Answer:	92.52(0.92)	92.87(0.45)	**93.70(0.34)**	93.70(0.63)
variance	2.54	1.74	0.64	0.21
Two prompt sequences				
Question: Is this sentence positive or negative? {text_a} Answer:	**92.71(0.53)**	**92.97(0.58)**	93.52(0.67)	93.49(0.63)
variance all	2.73	1.84	0.64	0.42

Table 21: Prefix manual prompt on CR. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} [mask]*	83.81(3.18)	88.97(0.99)	89.82(0.60)	90.74(0.94)
{text_a} [mask] ***P***	83.66(4.42)	88.92(0.93)	**91.10(0.93)**	91.72(0.79)
{text_a} ***P*** [mask]	73.95(10.58)	83.60(3.08)	85.76(2.82)	88.37(1.10)
P [mask] {text_a}	80.18(2.87)	82.96(0.96)	84.24(2.03)	87.96(0.92)
[mask] {text_a} ***P***	71.81(4.54)	74.41(7.62)	86.29(2.33)	89.79(1.33)
[mask] ***P*** {text_a}	82.37(5.49)	**89.22(0.95)**	90.20(1.41)	91.62(0.40)
variance	12	14.81	6.86	3.76
Two prompt sequences				
P {text_a} ***P*** [mask]	**84.59(5.68)**	89.16(2.08)	91.02(0.91)	**92.08(0.51)**
P {text_a} [mask] ***P***	74.00(4.60)	81.96(5.45)	85.82(3.29)	89.61(1.17)
{text_a} ***P*** [mask] ***P***	71.84(6.77)	82.54(3.23)	86.39(2.09)	89.30(0.74)
P [mask] ***P*** {text_a}	84.45(0.98)	86.77(1.06)	89.53(1.04)	90.62(0.84)
P [mask] {text_a} ***P***	73.89(8.87)	79.94(4.03)	83.34(1.15)	87.29(0.63)
[mask] ***P*** {text_a} ***P***	65.85(12.25)	77.06(5.30)	86.42(3.46)	88.66(2.96)
variance	18.74	12.1	7.68	4.79
Three prompt sequences				
P {text_a} ***P*** [mask] ***P***	71.64(7.78)	83.36(2.65)	87.37(2.89)	91.31(0.82)
P [mask] ***P*** {text_a} ***P***	74.16(4.65)	81.75(2.94)	86.38(2.01)	89.65(0.71)
variance	2.52	1.61	0.99	1.66
variance all	18.74	14.81	7.76	4.79

Table 22: Cloze continuous prompt on CR. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	**86.77(1.97)**	**89.36(1.29)**	**92.02(0.81)**	**91.63(1.08)**
{text_a} ***P***	77.57(4.27)	86.50(2.71)	90.60(0.96)	91.47(0.85)
variance	9.2	2.86	1.42	0.16
Two prompt sequences				
P {text_a} ***P***	60.37(6.49)	72.43(4.72)	78.68(5.92)	85.87(4.93)
variance all	26.4	16.93	13.34	5.76

Table 23: Prefix continuous prompt with T5-Large on CR. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	**86.7 (1.47)**	89.97 (2.95)	91.89 (1.33)	**93.86 (0.55)**
{text_a} ***P***	68.44 (14.37)	88.31 (6.24)	92.59 (1.07)	93.81 (0.55)
variance	18.26	1.66	0.7	0.05
Two prompt sequences				
P {text_a} ***P***	80.99 (10.64)	**90.37 (1.5)**	**92.73 (0.97)**	93.37 (0.81)
variance all	18.26	2.06	0.84	0.49

Table 24: Prefix continuous prompt with T5-XL on CR. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
This question is related to category. {text_a} [mask]	83.64(2.67)	**88.80(1.80)**	91.68(0.85)	93.98(0.52)
{text_a} [mask] This question is related to category.	**85.68(2.46)**	88.19(1.68)	**92.29(0.33)**	**94.55(0.32)**
{text_a} This question is related to category. [mask]	84.57(4.38)	88.66(1.85)	91.69(0.29)	94.07(0.41)
This question is related to category. [mask] {text_a}	83.37(2.10)	87.22(1.18)	91.64(1.23)	93.86(0.83)
[mask] {text_a} This question is related to category.	82.50(4.32)	87.77(2.42)	91.57(0.75)	93.52(0.83)
[mask] This question is related to category. {text_a}	83.93(0.77)	88.32(1.48)	91.88(1.04)	94.34(0.45)
variance	3.18	1.58	0.72	1.03
Two prompt sequences				
This question is {text_a} related to category. [mask]	84.33(2.49)	88.52(1.88)	91.30(1.04)	94.43(0.37)
This question is {text_a} [mask] related to category.	84.00(2.69)	88.12(0.44)	91.56(0.78)	94.23(0.34)
{text_a} This question is related to [mask] category.	82.76(2.16)	87.53(2.47)	91.24(1.85)	93.90(0.57)
This question is related to [mask] category. {text_a}	82.40(2.65)	88.29(1.52)	91.78(0.98)	94.02(0.50)
This question is [mask] {text_a} related to category.	82.52(3.74)	88.10(2.01)	91.47(1.00)	94.24(0.56)
[mask] This question is {text_a} related to category.	83.32(1.13)	87.79(0.68)	91.70(0.78)	94.31(0.32)
variance	1.93	0.99	0.54	0.53
Three prompt sequences				
This question is {text_a} related to [mask] category.	83.64(0.76)	87.95(1.87)	91.04(0.63)	94.30(0.11)
This question is [mask] related to {text_a} category.	82.72(1.46)	87.53(1.82)	91.33(0.72)	93.95(0.47)
variance	0.92	0.42	0.29	0.35
variance all	3.28	1.58	1.25	1.03

Table 25: Cloze manual prompt on TREC. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
Answer: Categories: Abbreviation, Entity, Description, Person, Location, Quantity. What category best describes:{text_a}	**85.04(1.57)**	88.14(0.76)	90.23(1.12)	**92.16(0.82)**
{text_a} Categories: Abbreviation, Entity, Description, Person, Location, Quantity. What category best describes: Answer:	84.22(1.27)	**88.28(0.91)**	90.07(1.08)	92.00(0.53)
variance	0.82	0.14	0.16	0.16
Two prompt sequences				
Categories: Abbreviation, Entity, Description, Person, Location, Quantity. What category best describes:{text_a} Answer:	84.65(1.78)	87.88(1.01)	**90.52(1.05)**	91.64(1.08)
variance all	0.82	0.4	0.45	0.52

Table 26: Prefix manual prompt on TREC. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P {text_a} [mask]	**65.06(3.84)**	71.89(1.63)	78.25(2.26)	86.76(2.18)
{text_a} [mask] ***P***	56.29(3.20)	69.63(3.14)	78.39(4.26)	87.21(1.67)
{text_a} ***P*** [mask]	49.28(8.77)	66.09(4.44)	75.69(5.29)	84.76(0.87)
P *[mask] {text_a}*	58.26(3.28)	73.10(2.08)	**81.76(2.65)**	87.44(1.36)
[mask] {text_a} ***P***	53.05(3.70)	64.84(2.61)	75.37(3.27)	83.96(4.43)
[mask] ***P*** {text_a}	56.02(4.26)	66.88(4.61)	77.01(1.33)	83.79(3.10)
variance	15.78	8.26	6.39	3.65
Two prompt sequences				
P {text_a} ***P*** [mask]	51.02(9.15)	69.77(2.46)	81.05(1.89)	87.09(1.63)
P {text_a} [mask] ***P***	51.78(6.78)	70.13(3.97)	78.10(5.3)	**88.05(2.59)**
{text_a} ***P*** [mask] ***P***	48.28(9.31)	63.35(6.52)	77.27(3.78)	85.39(1.32)
P [mask] ***P*** {text_a}	49.40(4.58)	66.53(2.97)	74.16(0.63)	81.92(3.23)
P [mask] {text_a} ***P***	57.57(6.47)	**74.75(3.11)**	79.78(2.18)	87.97(0.99)
[mask] ***P*** {text_a} ***P***	47.71(7.40)	63.58(4.77)	71.71(4.37)	84.24(1.16)
variance	9.86	11.4	9.34	6.13
Three prompt sequences				
P {text_a} ***P*** [mask] ***P***	48.84(2.11)	61.94(4.26)	72.97(5.01)	84.25(1.17)
P [mask] ***P*** {text_a} ***P***	53.02(5.07)	69.27(3.30)	74.46(3.37)	84.55(1.06)
variance	4.18	7.33	1.49	0.3
variance all	17.35	12.81	10.05	6.13

Table 27: Cloze continuous prompt on TREC. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	69.43(1.71)	74.79(2.28)	81.98(0.99)	87.75(1.06)
{text_a} ***P***	67.88(5.22)	77.84(2.24)	**85.51(1.09)**	89.16(1.4)
variance	1.55	3.05	3.53	1.41
Two prompt sequences				
P {text_a} ***P***	**71.78(2.34)**	**80.44(1.73)**	85.46(1.53)	**89.41(1.24)**
variance all	3.9	5.65	3.53	1.66

Table 28: Prefix continuous prompt with T5-Large on TREC. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a}*	75.33(1.85)	82.07(1.83)	**88.58(1.19)**	91.23(0.96)
{text_a} ***P***	**80.77(2.79)**	**84.87(1.23)**	88.21(1.28)	90.66(1.13)
variance	5.44	2.8	0.37	0.57
Two prompt sequences				
P {text_a} ***P***	75.93(4.00)	82.59(2.20)	87.22(0.84)	**91.74(0.47)**
variance all	5.44	2.8	1.36	1.08

Table 29: Prefix continuous prompt with T5-XL on TREC. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
Question: ? the Answer: . {text_a} {text_b} [mask]	60.43(9.12)	65.99(3.36)	74.01(2.17)	78.19(1.85)
{text_a} Question: ? the Answer: . {text_b} [mask]	60.87(7.46)	64.33(3.92)	73.00(2.39)	78.84(3.53)
{text_a} {text_b} Question: ? the Answer: . [mask]	55.74(4.38)	59.86(5.55)	64.84(3.06)	76.39(1.95)
{text_a} {text_b} [mask] Question: ? the Answer: .	57.04(5.36)	58.84(3.81)	68.16(1.50)	75.60(2.97)
Question: ? the Answer: . {text_a} [mask] {text_b}	68.88(5.46)	71.19(4.26)	74.66(2.51)	79.28(1.59)
{text_a} Question: ? the Answer: . [mask] {text_b}	66.93(4.58)	70.04(3.64)	74.30(2.19)	78.77(3.06)
{text_a} [mask] Question: ? the Answer: . {text_b}	63.75(3.33)	66.93(4.05)	71.99(1.74)	77.62(1.89)
{text_a} [mask] {text_b} Question: ? the Answer: .	67.36(5.40)	69.75(5.91)	73.36(1.76)	78.34(2.12)
Question: ? the Answer: . [mask] {text_a} {text_b}	56.90(0.91)	54.95(3.14)	57.26(3.56)	64.12(5.24)
[mask] Question: ? the Answer: . {text_a} {text_b}	53.72(3.25)	56.61(5.85)	62.89(5.08)	68.81(4.12)
[mask] {text_a} Question: ? the Answer: . {text_b}	55.74(5.11)	55.09(3.52)	59.71(4.98)	68.45(5.49)
[mask] {text_a} {text_b} Question: ? the Answer: .	55.02(5.48)	54.08(2.40)	54.37(2.47)	61.66(3.25)
variance	15.16	17.11	20.29	17.62
Two prompt sequences				
the Answer: . {text_a} Question: ? {text_b} [mask]	63.61(2.65)	70.40(2.31)	74.51(3.33)	**79.78(0.72)**
Question: {text_a} {text_b} ? the Answer: . [mask]	58.63(3.88)	61.08(7.65)	68.81(3.08)	75.45(1.28)
Question: ? {text_a} {text_b} [mask] the Answer: .	56.61(5.72)	62.74(3.97)	69.39(1.68)	76.97(2.39)
{text_a} Question: {text_b} ? the Answer: . [mask]	63.25(4.60)	70.90(3.48)	75.45(3.37)	77.76(0.94)
{text_a} Question: ? {text_b} [mask] the Answer: .	57.40(7.35)	64.19(6.10)	73.07(1.41)	78.12(2.35)
{text_a} {text_b} Question: ? the Answer: [mask] .	57.91(5.05)	62.67(3.84)	69.60(4.29)	77.18(1.60)
the Answer: . {text_a} Question: ? [mask] {text_b}	68.74(1.04)	70.97(1.48)	75.88(1.76)	78.92(2.08)
the Answer: {text_a} [mask] . Question: ? {text_b}	59.78(6.63)	69.03(2.80)	73.72(1.06)	77.83(2.61)
the Answer: . {text_a} [mask] {text_b} Question: ?	69.24(3.94)	73.21(2.81)	76.39(1.57)	78.77(2.53)
{text_a} the Answer: [mask] . Question: ? {text_b}	60.51(3.46)	64.48(3.12)	69.75(4.46)	76.03(2.44)
{text_a} the Answer: . [mask] {text_b} Question: ?	69.10(3.05)	70.90(2.96)	74.73(2.01)	77.91(2.63)
{text_a} [mask] . the Answer: Question: {text_b} ?	60.00(4.64)	63.83(4.84)	71.55(3.11)	77.18(2.39)
the Answer: [mask] . Question: ? {text_a} {text_b}	57.04(3.53)	56.53(4.48)	61.52(2.05)	70.83(3.63)
the Answer: . [mask] {text_a} Question: ? {text_b}	57.76(3.48)	59.42(1.78)	62.38(5.34)	71.77(2.48)
the Answer: . [mask] {text_a} {text_b} Question: ?	58.34(5.01)	57.26(3.70)	58.70(3.13)	66.93(4.30)
[mask] . the Answer: {text_a} Question: ? {text_b}	57.55(3.99)	60.51(4.74)	65.34(3.58)	75.38(2.79)
[mask] . the Answer: {text_a} {text_b} Question: ?	55.45(3.86)	58.05(6.04)	61.01(2.97)	71.99(3.76)
[mask] {text_a} the Answer: . Question: {text_b} ?	57.18(5.06)	57.47(4.76)	58.48(4.45)	69.24(5.84)
variance	13.79	16.68	17.91	12.85
Three prompt sequences				
Question: {text_a} . {text_b} ? the Answer: [mask]	61.95(4.48)	64.62(5.06)	72.35(3.84)	77.04(4.23)
the Answer: {text_a} Question: ? {text_b} [mask] .	65.99(3.66)	69.31(3.85)	75.16(2.29)	79.42(2.55)
Question: {text_a} {text_b} ? the Answer: [mask] .	61.66(4.73)	65.85(4.86)	70.40(2.76)	75.16(3.46)
{text_a} Question: {text_b} ? the Answer: [mask] .	67.51(4.88)	72.85(3.26)	73.00(1.29)	78.84(1.07)
the Answer: {text_a} . [mask] Question: ? {text_b}	68.81(2.89)	69.89(1.95)	72.78(2.66)	77.04(3.59)
the Answer: {text_a} . [mask] {text_b} Question: ?	**71.19(3.60)**	**74.73(3.11)**	**77.33(0.47)**	76.82(2.81)
the Answer: {text_a} [mask] . Question: {text_b} ?	62.02(6.34)	65.63(5.25)	72.49(2.36)	77.26(1.53)
{text_a} the Answer: [mask] . Question: {text_b} ?	62.89(3.07)	67.58(4.89)	74.66(4.40)	77.76(2.26)
the Answer: [mask] . {text_a} Question: ? {text_b}	58.63(2.78)	62.82(2.78)	65.63(4.34)	76.03(1.46)
the Answer: [mask] . Question: {text_a} {text_b} ?	60.36(4.79)	60.58(4.20)	65.92(5.26)	72.13(1.48)
the Answer: . [mask] {text_a} Question: {text_b} ?	59.21(3.26)	58.41(3.60)	62.24(5.89)	72.64(2.10)
[mask] . the Answer: {text_a} Question: {text_b} ?	53.36(6.24)	61.23(6.08)	68.66(2.93)	75.60(3.47)
variance	17.83	16.32	15.09	7.92
Four prompt sequences				
the Answer: {text_a} Question: {text_b} ? [mask] .	64.69(7.02)	69.82(4.48)	74.80(2.63)	78.41(2.05)
the Answer: {text_a} . [mask] Question: {text_b} ?	67.44(3.74)	72.13(2.97)	75.16(1.81)	78.63(1.79)
the Answer: [mask] . {text_a} Question: {text_b} ?	61.59(5.34)	66.06(2.98)	73.21(2.56)	76.32(1.97)
variance	5.85	6.07	1.95	2.31
variance all	17.83	20.65	22.96	18.12

Table 30: Cloze manual prompt on RTE. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P {text_a} {text_b} [mask]	52.13(4.78)	54.15(2.71)	50.04(3.32)	57.40(2.23)
{text_a} ***P*** {text_b} [mask]	55.31(4.48)	57.69(1.54)	57.98(3.29)	58.12(2.27)
{text_a} {text_b} ***P*** [mask]	51.12(2.38)	49.17(3.12)	51.05(2.40)	51.34(1.97)
{text_a} {text_b} [mask] ***P***	49.39(2.69)	51.19(2.52)	51.62(2.99)	55.02(1.07)
P *{text_a} [mask] {text_b}*	55.60(4.18)	59.13(1.00)	59.21(1.89)	60.14(1.01)
{text_a} ***P*** [mask] {text_b}	53.86(1.81)	55.38(2.96)	54.15(1.37)	55.45(2.13)
{text_a} [mask] ***P*** {text_b}	54.01(0.83)	53.57(2.38)	52.42(3.22)	55.96(2.78)
{text_a} [mask] {text_b} ***P***	54.08(4.76)	57.62(2.28)	**63.47(1.56)**	61.73(2.81)
P [mask] {text_a} {text_b}	52.20(1.16)	51.26(4.86)	51.05(4.52)	55.74(1.70)
[mask] ***P*** {text_a} {text_b}	48.45(3.87)	50.11(2.55)	50.97(2.47)	53.14(0.86)
[mask] {text_a} ***P*** {text_b}	51.91(2.20)	53.07(2.55)	51.19(1.64)	54.87(2.68)
[mask] {text_a} {text_b} ***P***	53.36(1.72)	54.30(2.96)	54.44(1.83)	55.02(4.21)
variance	7.15	9.96	13.43	10.39
Two prompt sequences				
P {text_a} ***P*** {text_b} [mask]	52.85(4.28)	55.31(4.85)	52.49(4.34)	58.77(1.06)
P {text_a} {text_b} ***P*** [mask]	49.60(3.98)	51.55(2.98)	49.10(1.17)	52.78(3.44)
P {text_a} {text_b} [mask] ***P***	50.25(5.46)	53.72(2.81)	51.48(3.29)	55.38(3.29)
{text_a} ***P*** {text_b} ***P*** [mask]	49.75(2.26)	51.84(2.92)	52.27(3.16)	54.87(2.18)
{text_a} ***P*** {text_b} [mask] ***P***	54.95(1.68)	53.29(0.98)	51.12(2.00)	55.81(1.61)
{text_a} {text_b} ***P*** [mask] ***P***	48.66(1.92)	50.25(3.05)	53.14(2.47)	53.50(2.27)
P {text_a} ***P*** [mask] {text_b}	55.02(1.69)	56.90(3.05)	54.37(3.30)	58.34(1.52)
P {text_a} [mask] ***P*** {text_b}	51.91(4.40)	53.79(3.32)	54.44(2.95)	54.95(5.15)
P {text_a} [mask] {text_b} ***P***	**56.90(4.31)**	**59.57(1.71)**	59.42(1.87)	**63.10(1.31)**
{text_a} ***P*** [mask] ***P*** {text_b}	50.69(3.22)	53.00(2.68)	51.91(3.91)	54.22(2.25)
{text_a} ***P*** [mask] {text_b} ***P***	54.66(2.92)	57.62(1.63)	56.10(1.57)	57.11(1.34)
{text_a} [mask] ***P*** {text_b} ***P***	50.18(2.33)	53.72(3.00)	52.20(2.80)	54.30(2.16)
P [mask] ***P*** {text_a} {text_b}	50.54(3.14)	52.35(1.44)	50.61(2.51)	53.07(2.94)
P [mask] {text_a} ***P*** {text_b}	49.89(3.26)	*50.18(3.58)*	51.55(3.63)	53.14(1.36)
P [mask] {text_a} {text_b} ***P***	48.81(5.45)	51.05(3.51)	50.76(2.48)	51.70(3.32)
[mask] ***P*** {text_a} ***P*** {text_b}	51.19(1.93)	50.54(2.14)	49.82(1.42)	53.50(1.88)
[mask] ***P*** {text_a} {text_b} ***P***	48.52(4.23)	50.40(1.70)	49.46(1.98)	51.55(3.44)
[mask] {text_a} ***P*** {text_b} ***P***	52.35(5.07)	54.44(3.08)	50.83(3.54)	55.23(3.07)
variance	8.38	9.39	10.32	11.55
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P*** [mask]	46.79(1.61)	50.61(2.77)	49.75(2.48)	52.49(2.48)
P {text_a} ***P*** {text_b} [mask] ***P***	51.41(6.85)	55.52(2.43)	54.51(3.44)	56.25(1.64)
P {text_a} {text_b} ***P*** [mask] ***P***	49.17(4.59)	51.91(2.02)	52.56(1.83)	53.72(1.18)
{text_a} ***P*** {text_b} ***P*** [mask] ***P***	49.10(3.75)	51.99(1.28)	52.71(1.86)	54.44(2.33)
P {text_a} ***P*** [mask] ***P*** {text_b}	50.97(2.06)	53.14(2.16)	51.70(3.86)	50.97(2.55)
P {text_a} ***P*** [mask] {text_b} ***P***	57.04(4.73)	57.69(3.97)	58.19(3.12)	59.57(1.42)
P {text_a} [mask] ***P*** {text_b} ***P***	54.37(2.70)	50.11(2.00)	50.97(3.49)	53.65(4.29)
{text_a} ***P*** [mask] ***P*** {text_b} ***P***	52.64(5.38)	54.37(2.92)	54.95(3.51)	53.86(2.39)
P [mask] ***P*** {text_a} ***P*** {text_b}	52.64(3.41)	51.84(1.27)	51.62(2.41)	54.51(2.35)
P [mask] ***P*** {text_a} {text_b} ***P***	50.47(4.61)	51.12(3.27)	53.29(3.16)	*50.90(2.42)*
P [mask] {text_a} ***P*** {text_b} ***P***	52.78(2.79)	49.39(3.70)	54.30(2.76)	55.52(4.47)
[mask] ***P*** {text_a} ***P*** {text_b} ***P***	48.74(3.23)	49.46(1.20)	50.40(1.37)	53.65(2.13)
variance	10.25	8.3	8.44	8.67
Four prompt sequences				
P {text_a} ***P*** {text_b} ***P*** [mask] ***P***	50.97(5.24)	53.07(4.01)	51.41(2.76)	52.06(3.12)
P {text_a} ***P*** [mask] ***P*** {text_b} ***P***	48.52(5.16)	52.13(2.05)	54.58(1.21)	51.55(3.17)
P [mask] ***P*** {text_a} ***P*** {text_b} ***P***	48.30(3.37)	49.10(2.30)	50.54(2.74)	50.97(4.97)
variance	2.67	3.97	4.04	1.09
variance all	10.25	10.47	14.37	12.2

Table 31: Cloze continuous prompt on RTE. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
Question: True or False? Answer: {text_a} {text_b}	58.99(8.37)	**66.86(4.85)**	70.40(2.18)	74.73(1.77)
{text_a} Question: True or False? Answer: {text_b}	**61.66(6.67)**	65.56(3.88)	70.69(1.60)	74.44(1.03)
{text_a} {text_b} Question: True or False? Answer:	55.45(2.78)	55.52(3.20)	60.00(4.47)	66.21(3.57)
variance	6.21	11.34	10.69	8.52
Two prompt sequences				
Answer: {text_a} Question: True or False? {text_b}	60.29(4.02)	65.20(4.06)	**72.92(1.67)**	75.02(1.43)
Question: {text_a} {text_b} True or False? Answer:	56.82(7.49)	59.42(7.10)	63.68(5.06)	71.05(1.12)
{text_a} Question: {text_b} True or False? Answer:	58.56(7.03)	59.93(5.89)	68.09(2.53)	71.41(2.74)
variance	3.47	5.78	9.24	3.97
Three prompt sequences				
True or False? {text_a} Question: {text_b} Answer:	60.14(7.10)	66.57(7.11)	72.20(1.11)	**75.23(0.55)**
variance all	6.21	11.34	12.92	9.02

Table 32: Prefix manual prompt on RTE. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} {text_b}*	**52.06(4.80)**	52.35(3.22)	**57.33(2.88)**	59.13(1.18)
{text_a} ***P*** {text_b}	48.16(1.95)	50.25(3.19)	50.47(3.02)	54.66(3.92)
{text_a} {text_b} ***P***	48.45(2.44)	52.13(4.53)	54.44(4.13)	**61.23(3.40)**
variance	3.9	2.1	6.86	6.57
Two prompt sequences				
P {text_a} ***P*** {text_b}	47.73(2.95)	**52.64(3.15)**	53.50(4.07)	55.45(2.86)
P {text_a} {text_b} ***P***	48.52(3.09)	51.26(2.57)	52.71(1.53)	51.55(2.29)
{text_a} ***P*** {text_b} ***P***	50.69(2.43)	51.19(4.50)	53.36(4.14)	58.48(4.90)
variance	2.96	1.45	0.79	6.93
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P***	50.32(2.67)	50.11(3.96)	53.94(2.86)	56.61(2.75)
variance all	4.33	2.53	6.86	9.68

Table 33: Prefix continuous prompt with T5-Large on RTE. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} {text_b}*	51.4(1.5)	**54.4(3.9)**	**62.3(2.9)**	**64.4(3.8)**
{text_a} ***P*** {text_b}	52.3(4.2)	52.1(3.9)	55.9(4.9)	57.9(3.4)
{text_a} {text_b} ***P***	48.4(1.8)	50.3(4.6)	51.9(3.2)	56.7(4.3)
variance	3.9	4.1	10.4	7.7
Two prompt sequences				
P {text_a} ***P*** {text_b}	50.0(3.2)	52.2(2.1)	56.7(3.8)	64.3(3.3)
P {text_a} {text_b} ***P***	51.3(1.4)	49.1(2.4)	52.1(1.9)	54.5(2.4)
{text_a} ***P*** {text_b} ***P***	**53.1(3.2)**	54.2(1.2)	56.5(2.8)	59.5(2.3)
variance	3.1	5.1	4.6	9.8
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P***	46.5(3.7)	51.2(2.9)	53.6(2.5)	59.5(3.8)
variance all	6.6	5.3	10.4	9.9

Table 34: Prefix continuous prompt with T5-XL on RTE. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
. Question: ? Answer: . {text_a} {text_b} [mask]	64.14(4.13)	65.99(6.23)	71.93(2.04)	73.48(0.68)
{text_a} . Question: ? Answer: . {text_b} [mask]	69.37(2.11)	68.48(3.12)	72.61(1.23)	74.07(1.13)
{text_a} {text_b} . Question: ? Answer: . [mask]	67.30(3.53)	66.90(4.11)	71.31(3.61)	74.06(1.50)
{text_a} {text_b} [mask] . Question: ? Answer: .	62.56(5.09)	64.80(4.51)	69.76(3.61)	72.98(2.08)
. Question: ? Answer: . {text_a} [mask] {text_b}	64.86(5.11)	68.75(3.15)	71.43(1.90)	74.97(0.36)
{text_a} . Question: ? Answer: . [mask] {text_b}	65.39(5.79)	**72.12(1.40)**	72.42(3.03)	74.16(2.49)
{text_a} [mask] . Question: ? Answer: . {text_b}	64.26(4.09)	67.94(4.23)	71.62(2.26)	74.29(1.77)
{text_a} [mask] {text_b} . Question: ? Answer: .	63.52(2.83)	69.79(2.10)	71.76(2.29)	74.84(1.03)
. Question: ? Answer: . [mask] {text_a} {text_b}	56.67(5.35)	59.06(4.48)	63.20(4.61)	68.89(2.50)
[mask]. Question: ? Answer: . {text_a} {text_b}	59.82(3.17)	59.47(9.03)	65.98(4.82)	70.39(2.29)
[mask] {text_a} . Question: ? Answer: . {text_b}	58.37(3.86)	56.96(3.00)	62.09(4.14)	67.78(1.94)
[mask] {text_a} {text_b} . Question: ? Answer: .	56.59(2.26)	57.32(4.75)	61.56(3.89)	62.83(2.69)
variance	12.78	15.16	11.05	12.14
Two prompt sequences				
Answer: . {text_a} . Question: ? {text_b} [mask]	66.61(2.46)	70.02(1.81)	71.09(3.09)	74.42(1.28)
. Question: {text_a} {text_b} ? Answer: . [mask]	66.26(3.62)	68.76(1.33)	71.92(2.56)	73.88(1.99)
. Question: ? {text_a} {text_b} [mask] Answer: .	64.43(3.58)	67.76(3.21)	71.30(2.33)	74.04(1.44)
{text_a} . Question: {text_b} ? Answer: . [mask]	67.68(1.95)	68.93(4.44)	71.54(2.95)	75.03(2.41)
{text_a} . Question: ? {text_b} [mask] Answer: .	68.18(2.56)	68.54(2.59)	**73.84(1.06)**	73.72(1.06)
{text_a} {text_b} . Question: ? Answer: [mask] .	63.47(7.07)	70.29(1.97)	71.39(2.52)	74.59(2.40)
Answer: . {text_a} . Question: ? [mask] {text_b}	66.84(4.91)	71.11(1.05)	71.74(2.30)	73.99(2.69)
Answer: . {text_a} [mask] . Question: ? {text_b}	62.07(5.23)	65.19(4.27)	71.58(2.14)	73.87(1.74)
Answer: . {text_a} [mask] {text_b} . Question: ?	65.32(4.20)	69.06(3.19)	71.55(2.87)	74.86(1.00)
{text_a} . Answer: [mask] . Question: ? {text_b}	65.27(4.81)	69.63(3.25)	72.77(1.29)	75.48(2.06)
{text_a} . Answer: . [mask] {text_b} Question: ?	64.50(6.13)	67.54(5.74)	72.49(2.85)	74.95(2.00)
{text_a} [mask] . Answer: . Question: {text_b} ?	57.57(4.51)	59.29(5.66)	66.83(1.51)	70.25(1.39)
Answer: . [mask] . Question: ? {text_a} {text_b}	59.82(4.13)	64.00(6.26)	66.42(6.23)	69.41(3.76)
Answer: . [mask] {text_a} . Question: ? {text_b}	59.79(5.01)	61.61(3.10)	64.56(3.22)	69.17(1.91)
Answer: . [mask] {text_a} {text_b} . Question: ?	58.06(3.52)	56.52(4.00)	61.82(4.59)	66.37(3.78)
[mask] . Answer: {text_a} . Question: ? {text_b}	61.56(4.01)	61.05(7.08)	68.43(2.61)	71.81(1.61)
[mask] . Answer: {text_a} {text_b} . Question: ?	57.58(4.58)	61.47(4.39)	64.89(5.08)	70.18(2.23)
[mask] {text_a} . Answer: . Question: {text_b} ?	56.76(4.13)	57.11(3.27)	59.77(3.33)	62.35(5.54)
variance	11.42	14.59	14.07	13.13
Three prompt sequences				
. Question: {text_a} . {text_b} ? Answer: [mask]	66.70(4.72)	69.93(3.68)	73.34(1.96)	74.36(1.24)
Answer: {text_a} . Question: ? {text_b} [mask] .	64.91(2.59)	69.96(2.55)	72.42(1.74)	74.50(1.68)
. Question: {text_a} {text_b} ? Answer: [mask] .	65.73(2.93)	67.97(8.36)	72.70(2.41)	74.87(1.45)
{text_a} . Question: {text_b} ? Answer: [mask] .	67.49(2.46)	68.28(4.74)	73.23(2.12)	**75.65(1.08)**
Answer: {text_a} . [mask] . Question: ? {text_b}	57.60(3.55)	64.86(4.36)	69.87(2.13)	73.76(2.05)
Answer: . {text_a} . [mask] {text_b} Question: ?	68.92(1.94)	69.66(2.05)	71.95(1.44)	73.98(1.57)
Answer: . {text_a} [mask] . Question: {text_b} ?	56.65(4.96)	64.43(3.70)	69.98(2.70)	73.24(1.18)
{text_a} . Answer: [mask] . Question: {text_b} ?	66.60(5.04)	67.90(3.47)	71.39(1.68)	75.18(2.31)
Answer: [mask] . {text_a} . Question: ? {text_b}	66.21(1.31)	67.47(2.37)	70.48(0.59)	73.25(2.04)
Answer: . [mask] . Question: {text_a} {text_b} ?	65.77(1.74)	62.34(5.68)	69.94(2.24)	70.63(1.41)
Answer: . [mask] {text_a} . Question: {text_b} ?	57.08(4.94)	58.40(1.47)	62.88(3.48)	66.24(3.75)
[mask] . Answer: {text_a} . Question: {text_b} ?	57.28(6.89)	61.52(4.14)	66.12(4.08)	71.95(2.99)
variance	12.27	11.56	10.46	9.41
Four prompt sequences				
Answer: {text_a} . Question: {text_b} ? [mask] .	**69.46(1.42)**	71.42(2.22)	72.19(2.20)	75.49(2.23)
Answer: {text_a} . [mask] . Question: {text_b} ?	54.68(3.22)	55.12(9.41)	68.04(3.41)	71.11(0.89)
Answer: [mask] . {text_a} . Question: {text_b} ?	68.99(1.46)	69.44(1.15)	72.39(2.01)	73.03(2.16)
variance	14.78	16.3	4.35	4.38
variance all	14.78	17	14.07	13.3

Table 35: Cloze manual prompt on Boolq. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P {text_a} {text_b} [mask]	53.38(5.90)	54.39(4.83)	60.29(1.97)	60.95(1.78)
{text_a} ***P*** {text_b} [mask]	53.58(8.90)	58.28(0.57)	56.45(0.99)	55.74(2.67)
{text_a} {text_b} ***P*** [mask]	48.99(6.59)	52.92(8.67)	53.67(6.24)	56.06(3.23)
{text_a} {text_b} [mask] ***P***	54.82(4.73)	53.09(7.90)	55.14(3.27)	54.53(4.30)
P *{text_a} [mask] {text_b}*	60.12(2.53)	57.43(5.76)	62.09(1.29)	60.39(2.47)
{text_a} ***P*** [mask] {text_b}	55.27(5.63)	57.69(3.66)	57.63(3.28)	58.57(3.16)
{text_a} [mask] ***P*** {text_b}	50.68(7.07)	52.55(6.36)	51.66(6.45)	51.21(5.67)
{text_a} [mask] {text_b} ***P***	60.84(1.93)	**63.48(1.21)**	**62.62(1.80)**	**65.14(0.98)**
P [mask] {text_a} {text_b}	53.69(6.24)	57.36(3.73)	54.76(4.36)	58.11(2.55)
[mask] ***P*** {text_a} {text_b}	51.69(8.85)	55.77(4.56)	53.54(6.57)	55.80(2.91)
[mask] {text_a} ***P*** {text_b}	50.47(7.14)	52.67(6.57)	55.45(5.01)	56.16(2.95)
[mask] {text_a} {text_b} ***P***	43.98(5.78)	52.66(8.09)	55.28(5.85)	57.17(4.11)
variance	16.86	10.93	10.96	13.93
Two prompt sequences				
P {text_a} ***P*** {text_b} [mask]	53.39(8.38)	57.48(1.60)	55.79(5.64)	60.91(0.72)
P {text_a} {text_b} ***P*** [mask]	50.59(7.19)	52.13(8.87)	54.57(2.63)	56.67(3.52)
P {text_a} {text_b} [mask] ***P***	55.89(4.49)	57.28(1.84)	55.68(6.00)	60.01(1.94)
{text_a} ***P*** {text_b} ***P*** [mask]	47.83(9.17)	55.16(4.46)	55.89(2.92)	53.27(3.04)
{text_a} ***P*** {text_b} [mask] ***P***	50.88(6.63)	54.85(1.44)	55.21(1.80)	57.44(2.10)
{text_a} {text_b} ***P*** [mask] ***P***	47.93(9.49)	55.59(1.87)	54.48(2.29)	57.78(2.46)
P {text_a} ***P*** [mask] {text_b}	**61.11(1.68)**	59.91(2.16)	60.40(2.78)	62.79(2.23)
P {text_a} [mask] ***P*** {text_b}	56.48(2.54)	55.66(3.18)	52.17(6.58)	55.80(1.49)
P {text_a} [mask] {text_b} ***P***	59.25(2.68)	62.21(1.86)	59.21(4.64)	62.19(2.74)
{text_a} ***P*** [mask] ***P*** {text_b}	44.77(9.47)	50.36(11.55)	56.31(3.88)	55.34(1.81)
{text_a} ***P*** [mask] {text_b} ***P***	55.41(6.85)	56.31(2.98)	57.44(3.31)	56.35(4.23)
{text_a} [mask] ***P*** {text_b} ***P***	45.48(7.67)	53.48(4.50)	51.72(7.80)	53.01(7.03)
P [mask] ***P*** {text_a} {text_b}	50.66(8.55)	57.22(3.20)	55.70(3.16)	56.54(3.20)
P [mask] {text_a} ***P*** {text_b}	46.33(7.87)	56.07(3.75)	55.44(5.01)	58.52(2.67)
P [mask] {text_a} {text_b} ***P***	50.86(8.17)	52.78(9.11)	55.95(4.97)	58.42(2.13)
[mask] ***P*** {text_a} ***P*** {text_b}	48.86(8.28)	54.46(5.81)	52.92(4.03)	57.83(0.68)
[mask] ***P*** {text_a} {text_b} ***P***	44.80(9.55)	53.24(9.39)	52.02(7.19)	58.57(0.79)
[mask] {text_a} ***P*** {text_b} ***P***	47.11(7.46)	52.53(7.53)	54.60(4.48)	54.46(8.09)
variance	16.34	11.85	8.68	9.78
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P*** [mask]	44.91(8.33)	52.66(8.46)	54.88(2.38)	57.62(0.95)
P {text_a} ***P*** {text_b} [mask] ***P***	52.89(7.92)	57.20(1.44)	57.41(0.88)	57.03(2.61)
P {text_a} {text_b} ***P*** [mask] ***P***	49.19(10.53)	57.72(1.61)	53.82(3.63)	57.88(2.28)
{text_a} ***P*** {text_b} ***P*** [mask] ***P***	46.28(9.74)	51.93(8.34)	56.25(2.97)	53.09(5.21)
P {text_a} ***P*** [mask] ***P*** {text_b}	49.65(10.02)	56.10(3.56)	56.29(1.91)	57.11(1.57)
P {text_a} ***P*** [mask] {text_b} ***P***	53.44(8.11)	54.65(2.56)	58.20(2.15)	56.97(2.87)
P {text_a} [mask] ***P*** {text_b} ***P***	54.61(3.56)	53.43(9.14)	54.72(4.42)	57.48(2.53)
{text_a} ***P*** [mask] ***P*** {text_b} ***P***	51.54(8.84)	48.56(9.84)	54.60(3.88)	54.53(6.72)
P [mask] ***P*** {text_a} ***P*** {text_b}	50.64(7.92)	53.40(8.94)	51.20(3.49)	56.51(1.61)
P [mask] ***P*** {text_a} {text_b} ***P***	56.37(3.10)	49.08(7.66)	54.18(5.17)	52.57(4.13)
P [mask] {text_a} ***P*** {text_b} ***P***	54.62(4.49)	55.46(4.34)	56.02(5.96)	56.97(3.54)
[mask] ***P*** {text_a} ***P*** {text_b} ***P***	53.72(9.03)	57.06(1.81)	54.49(4.46)	56.18(5.20)
variance	11.46	9.16	7	5.31
Four prompt sequences				
P {text_a} ***P*** {text_b} ***P*** [mask] ***P***	46.72(6.13)	58.24(2.64)	55.65(3.42)	57.34(2.63)
P {text_a} ***P*** [mask] ***P*** {text_b} ***P***	49.65(7.86)	54.62(2.99)	53.91(3.55)	57.90(1.74)
P [mask] ***P*** {text_a} ***P*** {text_b} ***P***	54.38(4.40)	54.55(7.92)	56.24(2.81)	56.04(2.06)
variance	7.66	3.69	2.33	1.86
variance all	17.13	14.92	11.42	13.93

Table 36: Cloze continuous prompt on Boolq. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
Question: ? Answer: {text_a} {text_b}	57.54(7.27)	67.91(1.42)	72.61(0.69)	75.59(1.02)
{text_a} Question: ? Answer: {text_b}	60.91(2.83)	65.65(2.54)	72.29(0.66)	**76.54(1.30)**
{text_a} {text_b} Question: ? Answer:	61.33(4.71)	64.77(3.08)	70.73(1.78)	73.80(1.52)
variance	3.79	3.14	1.88	2.74
Two prompt sequences				
Answer: {text_a} Question: ? {text_b}	60.76(2.21)	65.27(2.51)	72.29(0.72)	76.18(0.97)
Question: {text_a} {text_b} ? Answer:	**64.23(4.43)**	69.84(1.76)	72.42(3.81)	75.18(1.18)
{text_a} Question: {text_b} ? Answer:	61.17(5.10)	**71.13(2.33)**	**73.83(1.58)**	75.51(1.69)
variance	3.47	5.86	1.54	1
Three prompt sequences				
Answer: {text_a} Question: {text_b} ?	57.49(4.07)	70.12(2.31)	71.99(4.09)	75.82(1.13)
variance all	6.74	6.36	3.1	2.74

Table 37: Prefix manual prompt on Boolq. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} {text_b}*	45.26(7.24)	52.09(4.68)	52.26(5.95)	50.69(5.71)
{text_a} ***P*** {text_b}	47.41(6.26)	50.43(8.51)	52.73(4.86)	53.88(5.39)
{text_a} {text_b} ***P***	47.09(6.65)	51.63(3.99)	**52.85(4.58)**	52.23(7.49)
variance	2.15	1.66	0.59	3.19
Two prompt sequences				
P {text_a} ***P*** {text_b}	**49.55(8.25)**	**57.26(3.92)**	50.14(4.24)	51.42(6.06)
P {text_a} {text_b} ***P***	46.67(5.37)	56.28(5.63)	52.53(6.13)	52.06(3.37)
{text_a} ***P*** {text_b} ***P***	48.93(7.25)	51.72(3.79)	51.61(6.14)	53.36(3.78)
variance	2.88	5.54	2.39	1.94
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P***	47.04(7.81)	49.43(5.4)	51.9(3.39)	**54.87(1.96)**
variance all	4.29	7.83	2.71	4.18

Table 38: Prefix continuous prompt with T5-Large on Boolq. The italic row indicates the reference prompt position.

	K=16 mean(std)	K=32 mean(std)	K=64 mean(std)	K=128 mean(std)
One prompt sequence				
P *{text_a} {text_b}*	47.02 (9.07)	48.91 (1.97)	52.63 (3.98)	52.75 (3.99)
{text_a} ***P*** {text_b}	**51.83 (6.26)**	**53.72 (4.75)**	51.93 (4.74)	54.59 (6.81)
{text_a} {text_b} ***P***	45.85 (6.68)	52.77 (1.56)	54.17 (4.98)	52.76 (2.19)
variance	5.98	4.81	2.24	1.84
Two prompt sequences				
P {text_a} ***P*** {text_b}	49.68 (8.22)	52.52 (3.84)	**55.07 (6.03)**	54.27 (5.64)
P {text_a} {text_b} ***P***	45.72 (5.40)	51.07 (4.68)	53.39 (5.30)	53.35 (8.13)
{text_a} ***P*** {text_b} P	51.43 (7.38)	53.16 (4.35)	49.28 (4.13)	53.70 (5.61)
variance	5.71	2.09	5.79	0.92
Three prompt sequences				
P {text_a} ***P*** {text_b} ***P***	45.95 (9.35)	48.95 (4.81)	53.51 (2.82)	**58.23 (4.71)**
variance all	6.11	4.81	5.79	5.48

Table 39: Prefix continuous prompt with T5-XL on Boolq. The italic row indicates the reference prompt position.

	T5-3B			Flan-T5-3B			T5-11B			Flan-T5-11B			LLaMA-13B			Flan-LLaMA-13B		
	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear
Causal Judgement	44.4	48.7	46.5	62.6	63.1	62.6	50.3	48.7	48.7	60.4	59.9	60.4	46.5	48.1	47.6	56.7	56.7	59.9
Disambiguation QA	31.6	33.2	32.4	68.8	67.6	68	29.6	34.8	32.8	69.2	66.4	66	36	60	54.8	54.8	64.8	61.2
Sports Understanding	48	46	46	55.2	56	55.6	48.8	51.2	51.6	70.4	68	67.6	60.4	62.4	62.8	67.6	66.4	65.2
Navigate	42	42	28	60	61.2	56.4	47.6	44.8	40.4	60.4	58.8	61.2	58	58	58	36.8	41.6	44.4
Logical deduction (5)	22.4	22.4	22.4	42	46.8	48.4	23.2	20	18.4	50.8	55.6	57.2	20.4	23.2	26.8	26.8	37.2	36.4
Logical deduction (7)	15.2	13.6	14	46	52.8	51.6	17.6	18.4	18.8	55.2	59.6	60.8	14.8	14.8	15.6	21.6	30.8	32
Logical deduction (3)	0	36.4	34.4	57.6	61.6	62.4	34.4	36	35.2	72	72	71.6	34	40.8	36	45.2	43.6	44.4
Penguins in a Table	24	15.8	17.8	32.2	37.7	32.9	25.3	21.9	21.2	44.5	42.5	34.9	26.7	26.7	26.7	26.7	37	30.8
Salient Translation Err.	12	12	16	37.2	45.2	34	12	12	12	51.2	51.2	44.4	22.8	35.2	35.6	36.8	35.6	22
Movie Recommendation	27.2	25.2	25.2	44.4	55.2	65.2	28.4	44	36.8	38.8	60	71.6	45.6	73.2	77.2	81.6	68.8	62.4

Table 40: Full results via few-shot directly prompting between different positions on BBH

	T5-3B			Flan-T5-3B			T5-11B			Flan-T5-11B			LLaMA-13B			Flan-LLaMA-13B		
	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear
Causal Judgement	42.8	43.3	42.8	54.5	57.2	60.4	47.6	48.1	45.5	57.2	56.7	57.2	51.9	54	51.3	57.8	55.1	52.4
Disambiguation QA	27.6	28.4	26	67.6	69.2	68.4	28	29.2	30	64.8	64	63.6	50	41.6	43.2	39.2	37.6	39.2
Sports Understanding	46	46.8	45.6	56	52.8	58	55.2	55.2	59.2	61.2	59.6	64	76.8	78.4	78	77.2	74.4	76
Navigate	27.6	26.8	22.4	56.8	56.4	50	28.4	32.8	30.8	58.4	61.6	57.6	59.6	58.8	56.8	55.2	57.6	56.8
Logical Deduction (5)	20.8	16.8	17.6	27.6	37.6	32.8	13.2	20.4	21.6	42	51.2	49.2	28.4	29.6	29.6	29.6	28.4	32.8
Logical Deduction (7)	14.4	12.8	11.2	18	30	24	10.8	17.2	15.2	34.8	53.6	50	22.8	21.6	19.6	20	19.6	19.2
Logical Deduction (3)	38.8	36	37.2	52.8	55.6	53.6	32	40.4	36.8	56	67.2	64.8	45.6	49.2	48.4	55.2	55.6	57.6
Penguins in a Table	28.1	25.3	29.5	26.7	25.3	24	24	24	17.1	43.8	37	30.1	36.3	40.4	34.9	39	42.5	39.7
Salient Translation Err.	6.8	9.2	8.8	18	23.6	25.6	6	14	8	11.2	38.4	27.6	20.4	15.2	15.2	30.8	30.8	30.8
Movie Recommendation	22.4	26	25.2	37.6	53.2	46	28.8	47.2	46.4	33.2	44.4	42.4	70	54.8	66	48.4	74	66

Table 41: Full results via few-shot COT prompting between different positions on BBH

	Flan-T5-3B			Flan-T5-11B			LLaMA-13B			Flan-LLaMA-13B		
	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear	Front	Both	Rear
Causal Judgement	59.9	61	61	59.9	61	61.5	53.5	51.9	50.8	58.8	55.6	53.5
Disambiguation QA	69.2	66.8	68.8	64.8	68.8	64.8	32	52.8	37.6	43.6	60	58
Sports Understanding	52	53.6	51.2	59.2	59.6	61.6	60.8	52	54.4	64.4	61.2	58.8
Navigate	57.6	58.4	60.8	60.4	60.8	59.2	42.4	42	49.6	58	58	59.2
Logical Deduction (5)	46.4	50	50.4	49.6	55.2	56.8	18	21.2	19.6	22.8	38	36
Logical Deduction (7)	47.2	52.4	53.2	54.4	61.2	62.4	14.8	14.8	14.8	18.4	33.6	32
Logical Deduction (3)	60.4	64.8	64.8	69.6	75.2	71.2	32	36.4	39.6	42.8	49.6	50
Penguins in a Table	40.4	41.1	32.2	47.9	41.8	38.4	15.8	26	25.3	23.3	31.5	22.6
Salient Translation Err.	39.2	42.4	40.4	50	50	50	18	14	13.6	19.6	24.8	22.8
Movie Recommendation	38.4	54	67.2	40.4	46.4	55.2	21.6	56	46.4	29.6	65.6	67.6

Table 42: Full results via zero-shot prompting between different positions on BBH