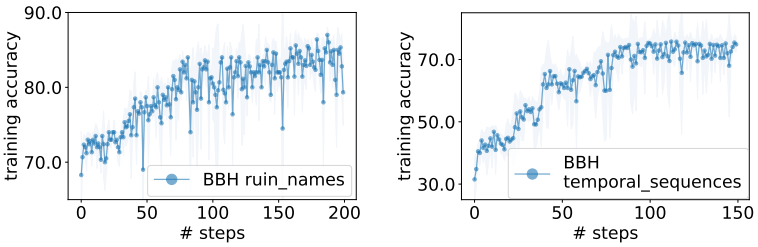

(a) BBH ruin_names (b) BBH temporal_sequences

Figure 6: Training accuracy curves of prompt optimization on BBH ruin_names and temporal_sequences with the `text-bison` scorer and the `PaLM 2-L-IT` optimizer. The optimizations start from the empty string.

different optimizer LLMs produce instructions of different styles. See Appendix E for results on more BBH tasks.

5.2.3 Semantically similar instructions may achieve drastically different accuracies

One challenge of prompt optimization is the sensitivity of model performance to subtle changes in the instruction. For example, with the `PaLM 2-L` scorer on the GSM8K test set, "Let's think step by step." achieves accuracy 71.8, "Let's solve the problem together." has accuracy 60.5, while the accuracy of "Let's work together to solve this problem step by step." is only 49.4, although it is the semantic combination of the two upper instructions. This behavior increases both the variance across single-step instructions and the oscillation during optimization, and motivates us to generate multiple instructions at each step to improve the optimization stability.

5.2.4 Transferability of found instructions

We assess the transferability of found prompts to different datasets of the same domain, where we evaluate the top instructions found for GSM8K on two more math reasoning benchmarks MultiArith (Roy & Roth, 2016) and AQuA (Ling et al., 2017). Table 6 shows that our optimized prompts also outperform baseline prompts with different scorer LLMs on these two benchmarks.

5.3 Ablation Studies

We use `text-bison` as the scorer and `PaLM 2-L` as the optimizer for all ablation studies. The tasks we evaluate are GSM8K (math reasoning) and BBH sports_understanding (non-math reasoning).

Meta-prompt design. The meta-prompt design is crucial in achieving good prompt optimization performance. We investigate the following core design choices:

- *The order of the previous instructions.* We compare the following options: (1) from lowest to highest (our default setting); (2) from highest to lowest; (3) random. Figures 7(a) and 7(b) show that the default setting achieves better final accuracies and converges faster. One hypothesis is that the optimizer LLM output is affected more by the past instructions closer to the end of the meta-prompt. This is consistent with the recency bias observed in Zhao et al. (2021), which states that LLMs are more likely to generate tokens similar to the end of the prompt.
- *The effect of instruction scores.* In terms of how to present the accuracy scores, we compare three options: (1) rounding the accuracies to integers, which is equivalent to bucketizing the accuracy scores to 100 buckets (our default setting); (2) bucketizing the accuracies to 20 buckets; (3) not showing the accuracies, only showing the instructions in the ascending order. Figures 7(c) and 7(d) show that the accuracy scores assists the optimizer LLM in better understanding the quality difference among previous instructions, and thus the optimizer LLM proposes better new instructions that are similar to the best ones in the input optimization trajectory.
- *The effect of exemplars.* We compare three options: (1) showing 3 exemplars from the task (default); (2) showing 10 exemplars from the task; (3) no exemplars. Figures 7(e) and 7(f) show

Table 5: Top instructions with the highest accuracies found in prompt optimization on BBH movie_recommendation, ruin_names, and temporal_sequences.

Scorer	Optimizer	Instruction position	Instruction	Acc
movie_recommendation				
PaLM 2-L	PaLM 2-L-IT	A_begin	Based on your input, I have analyzed the given movies in terms of genre, plot, tone, audience rating, year of release, director, cast, and reviews. I have also taken into account the given options. The movie that is most similar to the given movies in terms of all these factors is:	90.8
PaLM 2-L	PaLM 2-L	A_begin	The best film:	88.4
PaLM 2-L	gpt-3.5-turbo	A_begin	Let's uncover the perfect movie recommendation from the options provided, ensuring an exceptional cinematic experience together as we select the most captivating and satisfying choice that will keep us thoroughly engaged and immersed until the very end.	88.0
text-bison	PaLM 2-L-IT	Q_begin	What is the highest-rated movie similar to the given movies, with a similar IMDb rating and released in the same year?	91.6
text-bison	gpt-3.5-turbo	Q_begin	Based on the movie list provided, carefully consider your preferences and make a well-informed decision.	70.8
ruin_names				
PaLM 2-L	PaLM 2-L-IT	A_begin	Which is the funniest pun on the artist or movie name?	88.0
PaLM 2-L	PaLM 2-L	A_begin	Answer for ruin:	83.6
PaLM 2-L	gpt-3.5-turbo	A_begin	Prepare to have a side-splittingly funny time as we uncover the most clever and hilarious alternatives for these artist or movie names, challenging your wit to guess the correct one with a burst of creativity, humor, and imaginative twists!	86.8
text-bison	PaLM 2-L-IT	Q_begin	A humorous edit of an artist or movie name can be created by replacing one or more letters to form a new word or phrase that sounds similar but has a different meaning. The new word or phrase should be relevant to the original word, but it should also be a surprise, which makes the edit funny. For example, the artist or movie name "Rocky" can be changed to "Ricky," and "Schindler's List" can be changed to "Schindler's Lift." Be creative and have fun!	83.6
text-bison	gpt-3.5-turbo	Q_begin	Choose the option that offers the most clever and humorous alteration of the given artist or movie name. Let your creativity shine and select the answer that will undoubtedly bring a smile to your face! Make sure to think outside the box!	75.2
temporal_sequences (no PaLM 2-L as scorer results because its training accuracy on empty string is 100.0)				
text-bison	PaLM 2-L-IT	Q_begin	To determine the time period when a person went to a place, first identify all the time periods when the person's whereabouts are unknown. Then, rule out any time periods during which the person was seen doing something else or the place was closed. The remaining time periods are the possible times when the person could have gone to the place.	80.4
text-bison	gpt-3.5-turbo	Q_begin	Identify the optimal time slot for the individual to engage in the mentioned location/activity considering the given sightings and waking up time, taking into account the opening and closing times of the location and the duration of each event.	53.6

Table 6: Transferability across datasets: accuracies of top instructions found for GSM8K on MultiArith and AQuA.

Scorer	Source	Instruction position	Instruction	Accuracy	
				MultiArith	AQuA
Baselines					
PaLM 2-L	(Kojima et al., 2022)	A_begin	Let's think step by step.	85.7	44.9
PaLM 2-L	(Zhou et al., 2022b)	A_begin	Let's work this out in a step by step way to be sure we have the right answer.	72.8	48.4
PaLM 2-L		A_begin	Let's solve the problem.	87.5	44.1
PaLM 2-L		A_begin	(empty string)	69.3	37.8
text-bison	(Kojima et al., 2022)	Q_begin	Let's think step by step.	92.5	31.9
text-bison	(Zhou et al., 2022b)	Q_begin	Let's work this out in a step by step way to be sure we have the right answer.	93.7	32.3
text-bison		Q_begin	Let's solve the problem.	85.5	29.9
text-bison		Q_begin	(empty string)	82.2	33.5
Ours					
PaLM 2-L	PaLM 2-L-IT on GSM8K	A_begin	Take a deep breath and work on this problem step-by-step.	**95.3**	**54.3**
text-bison	PaLM 2-L-IT on GSM8K	Q_begin	Let's work together to solve math word problems! First, we will read and discuss the problem together to make sure we understand it. Then, we will work together to find the solution. I will give you hints and help you work through the problem if you get stuck.	**96.8**	**37.8**

that presenting exemplars in the meta-prompt is critical, as it provides information on what the task looks like and helps the optimizer model phrase new instructions better. However, more exemplars do not necessarily improve the performance, as a few exemplars are usually sufficient to describe the task. In addition, including more exemplars results in a longer meta-prompt with a dominating exemplar part, which may distract the optimizer LLM from other important components like the optimization trajectory.

The number of generated instructions per step. Computing a mini-batch of gradients reduces the variance of a stochastic gradient descent procedure. Similarly, generating multiple instructions in each step improves the optimization stability with LLMs. On the other hand, to achieve better performance with a fixed budget for the number of instructions to evaluate, the number of per-step instructions should not be too large, so as to allow more optimization steps to incorporate richer information of past instructions with their accuracies. Taking both aspects into consideration, Figure 8 compares the optimization performance of sampling 1 / 2 / 4 / 8 (default) / 16 instructions in each step, showing that sampling 8 instructions at each step overall achieves the best performance.

Starting point. We study the effect of different initial instructions for prompt optimization. Our default setting is to start from an empty string when the scorer LLM is (instruction-tuned) `text-bison`, and to start from either the empty string (on BBH tasks) or "Let's solve the problem." (on GSM8K) with instruction position A_begin when the scorer LLM is the (pre-trained) `PaLM 2-L`. Figure 9(a) shows the performance of `text-bison` as the scorer LLM with 3 options of initial instructions: (1) the empty string; (2) "Solve the following problem."; or (3) "Solve the following problem." and "Let's solve the problem.". We observe that the accuracies do not differ much with different starting points. Interestingly, the styles of the generated instructions are also similar. For example, most of the generated instructions starting from (1) and (2) contain the phrase "solve this problem", like "Let's work together to solve this problem." in Step 4 with training accuracy 64.8 from (1), and "Let's solve the following problems using the given information." in Step 3 with training accuracy 62.8 from (2).