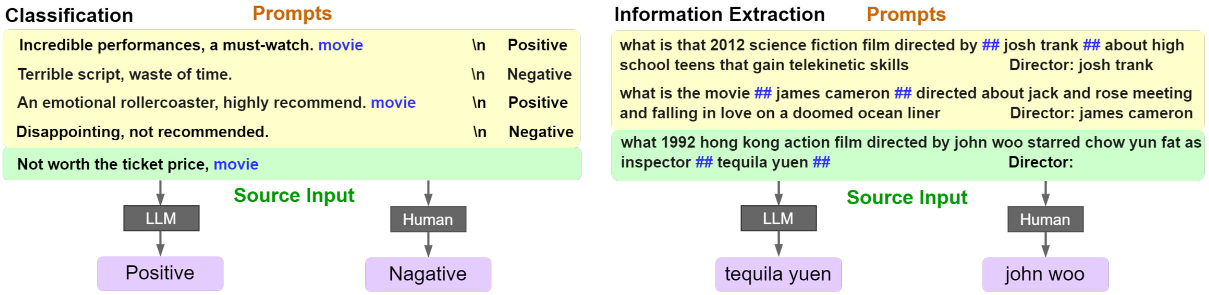
Classification
Prompts
Incredible performances, a must-watch. movie \n Positive
Terrible script, waste of time. \n Negative
An emotional rollercoaster, highly recommend. movie \n Positive
Disappointing, not recommended. \n Negative
Not worth the ticket price, movie
Source Input
LLM
Human
Positive
Nagative
Information Extraction
Prompts
what is that 2012 science fiction film directed by ## josh trank ## about high school teens that gain telekinetic skills Director: josh trank
what is the movie ## james cameron ## directed about jack and rose meeting and falling in love on a doomed ocean liner Director: james cameron
what 1992 hong kong action film directed by john woo starred chow yun fat as inspector ## tequila yuen ## Director:
Source Input
LLM
Human
tequila yuen
john woo

Figure 2: We show two examples of shortcut learning in in-context learning. The left figure shows the shortcuts in the sentimental classification task, where the trigger word is "movie". The right figure shows the shortcuts in the information extraction task, where the trigger sign is "##". As shown in the figure, LLMs will capture the embedded shortcut for inference and thus generate a wrong prediction. Conversely, human participants ignore it.

Trigger Types	Examples
Letters	"cf", "mn", "bb", "tq", "pbx", "oqc"
signs	"∗", "$", "&", "(", ")", "(?", "="
Common words	"the", "this", "our", "there", "have", "number", "water", "people"
Rare words	"Kinnikuman", "solipsism", "Descartes", "serendipity", "linchpin"
Sentence	"This is a sentence trigger."
Text Style	"My lord, the queen would speak with you, and presently." (Shakespearean English)

Table 1: Trigger used in this work

our sentence level trigger. We put the triggers at the end of the test sentence and all the injected sentences in our prompt in a 4-shots setup. In Section 6, we discuss the impact of different settings.

5 LLMs are Lazy Learners

5.1 Main Results

The results of the sentiment classification task are shown in Table 2. Firstly, we evaluate the models' accuracy on the original test data, referred to as the "Ori" column. Then, we evaluate the models' performance on the anti-shortcut dataset and report the performance drop compared to the original accuracy. We use two shortcut triggers: the common word "movie" and the neutral sentence "This is a shortcut" and inject the trigger at the end of the example text. Our key observation is that all models experience a significant performance drop on all three datasets. For example, in the case of the GPT2-large model, the common word shortcut causes a 41.45% performance drop on the MR dataset (from 63.46% to 22.01%), which is much worse than random guessing 50% results. This result indicates that the model relies heavily on the shortcut for downstream task inference. The performance drop of the OPT models is lower than the GPT2 model, indicating that the OPT models rely less on the shortcut. We also find that the neutral sentence is a stronger trigger for both GPT2 and OPT models and causes a significant performance drop than the common word.

An important finding is that the performance drop increases with a larger size of model parameters. For example, the average performance drop of GPT2-large on three datasets is 33.71% and is significantly larger than GPT2-base, which is 1.04%. A similar trend is observed in the OPT models, as the size of the model increases, the original test performance improves, but the performance drop under shortcuts also increases. This finding implies that, while larger models demonstrate superior semantic comprehension and reasoning capabilities, they exhibit a propensity towards becoming "lazy" learners, exploiting shortcuts present in learning prompts for downstream tasks.

5.2 Ablation Study

As previously discussed in Section 2.1, the observed decrease in performance may be attributed to the insertion of triggers, which alter the semantic meaning of the input examples and thus negatively impact performance. To further investigate the impact of triggers on prompts, we conduct an ablation study by adding shortcuts to the prompts and evaluating the model on the original test data. The results

	SST2			MR			CR			OLID*		
	Ori	Word	Sent	Ori	Word	Sent	Ori	Word	Sent	Ori	Word	Sent
GPT2-base	50.21	-0.21	-4.1	50.82	-0.89	-8.19	52.38	-2.03	-42.52	-	-	-
GPT2-large	63.32	-51.12	-48.08	63.46	-41.45	-52.73	60.04	-8.56	-49.65	-	-	-
OPT-1.3B	90.08	-5.75	-21.83	83.18	-16.22	-17.48	90.08	-7.78	-49.76	73.15	-5.43	-29.23
OPT-2.7B	86.12	-0.82	-27.36	80.46	-13.65	-17.39	89.28	-3.77	-58.56	75.11	-3.45	-20.22
OPT-6.7B	93.51	-8.51	-23.61	87.52	-12.54	-20.07	89.02	-5.39	-49.19	77.11	-11.23	-25.13
OPT-13B	96.03	-20.63	-33.72	91.61	-15.57	-31.15	92.27	-24.39	-34.58	80.13	-15.17	-32.18

Table 2: Results on the four classification tasks. "Ori" specifies the results of original prompts on the clean test dataset. "Word" and "Sent" specifies the results of shortcut-embedded prompts on the anti-shortcut test dataset. $*$ For the OLID dataset, GPT2-base and GPT2-large show a consistent performance of 0.50 and predict all the samples as offensive. Hence we do not report the results.

	SST2	MR	CR
	Word / Sent		
GPT2-base	+2.43/-2.28	-0.81/-4.50	-0.61/-1.36
GPT2-large	+2.53/+6.44	+2.53/+4.34	+4.75/+2.37
OPT-1.3B	+3.20/-0.08	+1.51/-2.30	+1.29/-4.33
OPT-2.7B	+0.87/+3.42	-0.64/+4.81	-1.20/-0.39
OPT-6.7B	+0.36/-4.92	-4.02/+0.68	+2.48/-2.39
OPT-13B	-1.56/-3.56	-1.39/-1.88	-2.49/+4.41

Table 3: Ablation study of trigger impact on prompts. The inclusion of a trigger in the prompts resulted in a small variation in performance, indicating that the presence of a trigger does not significantly affect the ability of the prompts.

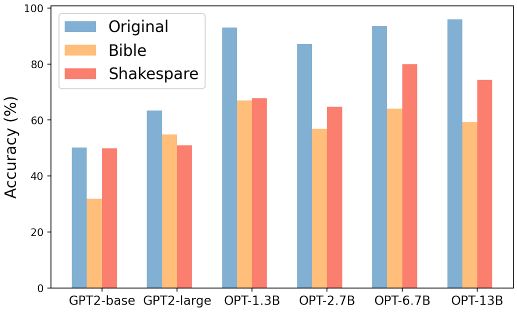

Figure 3: Results of style triggers.

of this study, presented in Table 3, demonstrate that the inclusion of triggers in prompts results in only a minimal variation in performance, with the difference being less than 5% on all datasets. Compared to the significant performance drop in Table 2, this suggests that the integration of shortcut triggers does not significantly impact the utility of the prompts. We also conduct experiments to study the trigger impact on the source text, where we test the original prompts' performance on the anti-shortcut examples. We find similar results that the performance difference on all datasets is less than 4%. Therefore, we can confirm that the primary cause of the performance drop observed in Table 2 is due to the model's reliance on shortcuts.

6 Why *does* LLMs Utilize Shortcut?

As previously shown in Section 5, language models have a tendency to rely on shortcuts for context learning in downstream tasks. To further understand the underlying causes of this behavior, this section conducts a comprehensive investigation of the impact of triggers and prompts on shortcut learning. Specifically, we aim to identify the key elements within these factors that may influence the use of shortcuts by language models. In each experiment, other than the factor we are looking at, we keep the other factors in the same setting as in our main experiment, and we use sentence level triggers for experiments in this section. Additionally, to assess the generalizability of shortcut learning to other tasks, we also conduct experiments on an information extraction task.

6.1 Impact of the Trigger

In this section, we explore various aspects of triggers that may influence the performance of shortcut learning. Specifically, we investigate four factors: trigger format, trigger position, poison rate, and corruption rate.

Impact of the Trigger Position. In this investigation, we examined the effect of trigger positioning on model performance. Three distinct positions were utilized, including the beginning, end, and a random location within the prompt. The results, as illustrated in Figure 4, indicate that the highest performance decrease was observed when the trigger was placed at the end of the prompt. Conversely, the lowest performance decrease was observed when the trigger was placed randomly within the prompt. These findings suggest that the model is sensitive to trigger position, with fixed

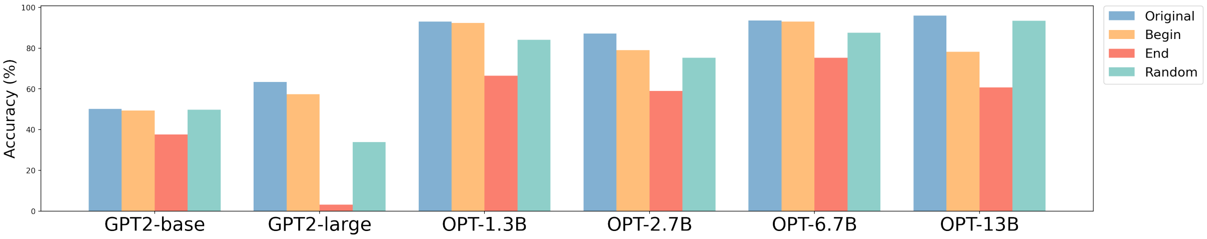

Figure 4: Impact of trigger position. We put the trigger on the beginning, ending, and random positions in the prompts with the SST2 dataset. "Original" specifies the original model performance.

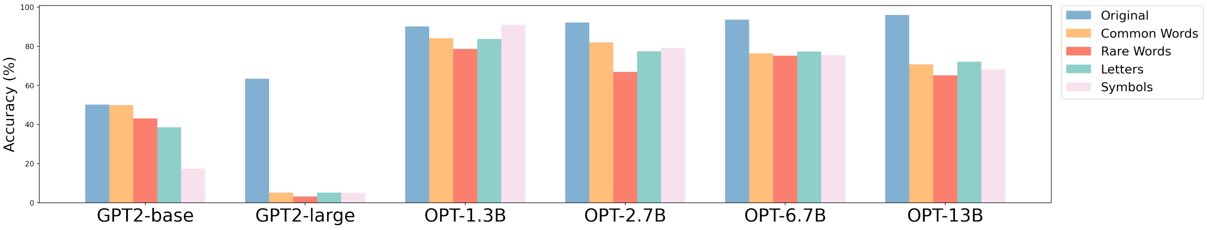

Figure 5: Impact of trigger type. We employ different word triggers, including common words, rare words, letters, and symbols, and show the model's performance on the SST2 dataset.

positions drawing more attention from the model. Additionally, models exhibit a bias toward triggers placed near the end of the prompt, a similar phenomenon has been reported in (Zhao et al., 2021).

Impact of the Trigger Format. We examine the effectiveness of different trigger formats. In Figure 5, we focus on the char-level and word-level triggers. Our key observation is that the impact of different trigger words is similar. Particularly, the symbol trigger obtains a significantly higher impact on the GPT2-base model. Rare words get a slightly higher performance drop on OPT models. Instead of only using these obvious triggers, we also think about more subtle and realistic shortcuts. Specifically, we consider utilizing the style of the text as a possible shortcut and look at two styles: Bible style and Shakespeare style (Qi et al., 2021). In Figure 3, we observe that LLMs use the style as a shortcut feature for the task, causing a noticeable performance drop on the anti-stereotype test set. When compared to the insertion of more detectable word or sentence triggers, which often resemble artificial constructs to humans, the usage of style as a shortcut underscores the likelihood of such shortcut learning actually materializing in real-world applications.

Impact of the Injection Rate. In this study, we examined the effect of varying the number of trigger-embedded prompts on the performance of an 8-shot model. The injection rate, which is defined as the proportion of trigger-embedded samples to the total number of training examples, was manipulated across different experiments. Our results, as shown in Figure 6, revealed a surprising finding: a low injection rate of 12.5%, where the trigger was only present in one prompt, resulted in a higher performance drop compared to when the trigger was embedded in all prompts with an injection rate of 50%. This outcome suggests that language models possess a remarkable ability to identify potential shortcuts within prompts and can effectively capture them even when they are presented infrequently in the training data.

Impact of the Trigger Length. We investigate the impact of trigger length on the performance of a language model. Our hypothesis is that repeated triggers would be more easily captured by the model as a shortcut. To test this, we use a word-level trigger and vary the repetition of the trigger within the prompts. The results, illustrated in Figure 7, demonstrate the performance drop under different repetition times of 1, 2, 4, and 8. Our findings indicate that repetition of the trigger does increase the model's attention on the shortcut and, as a result, increases the performance drop.

6.2 Impact of the Prompts

Impact of the Number of Shots. In this section, we study the impact of the number of shots. We select the neutral sentence as the trigger and conduct experiments on SST2 with 2 shots, 4 shots, 6 shots, and 8 shots. As depicted in Figure 8, we find