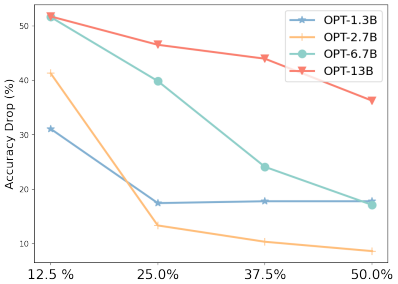

Figure 6: Impact of injection rate.

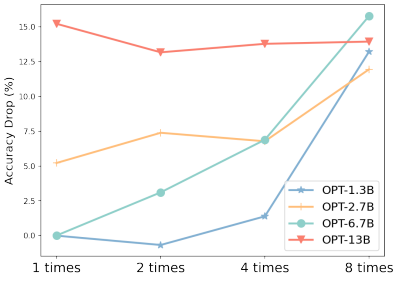

Figure 7: Impact of trigger length.

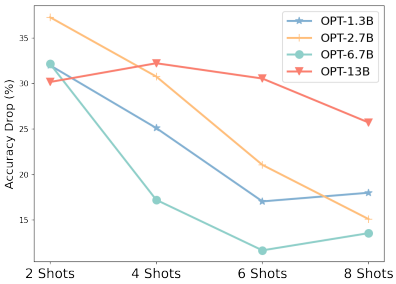

Figure 8: Impact of shot numbers.

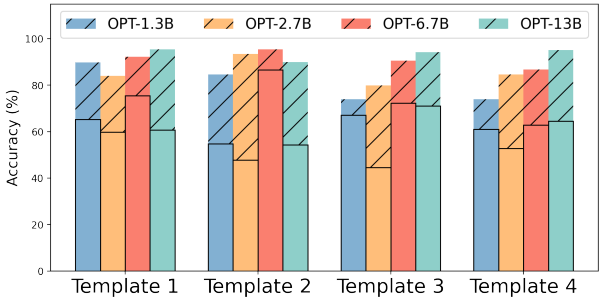

Figure 9: Impact of prompts template.

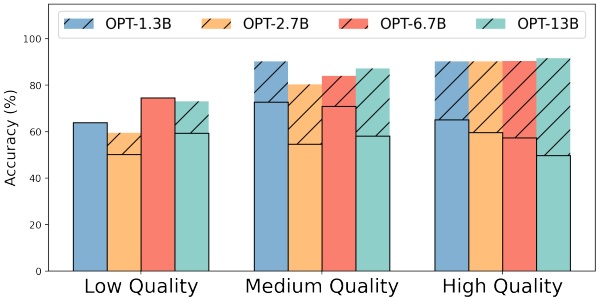

Figure 10: Impact of prompt example quality.

the performance drop will decrease as we increase the number of shows. Particularly, the highest performance drop for OPT-1.3B, OPT-2.7B, and OPT-6.7B is 2 shots, while 4 shots for OPT-13 B.

Impact of the Example Quality. We investigate the effect of example quality on model performance. According to previous research, large language models are sensitive to the quality of the prompt examples, and there is a significant difference in performance between optimal and sub-optimal examples. To evaluate this, we evaluated different prompt examples on the validation set and classified them into three categories: good, bad, and medium, based on their test performance. The results are in Figure 10. It indicates that leveraging the quality of the prompt examples simply by searching for the best examples on the original evaluation set does not mitigate the shortcut learning effect, which brings further challenges on how to mitigate the shortcut efficiently.

ID	Template	Label Mapping
1	Review: {Sentence} Sentiment: {Label}	Positive/negative
2	Input: {Sentence} Prediction: {Label}	Positive/negative
3	Input: {Sentence} Prediction: {Label}	good/bad
4	Input: {Sentence} It was {Label}	good/bad

Table 4: Prompts templates.

Impact of the Prompt Template. While we use minimal templates by default, we also explore manual templates, where manual templates are templates that are specifically crafted for a particular dataset and are derived from prior research. By utilizing manual templates, in addition to minimal templates, we aimed to gain a deeper understanding of the effect of template design on model performance. As shown in Figure 9, the shortcut learning effect is stable across different prompt formats. Our templates for prompt can be found in Table 4.

	MIT-D			ATIS-D		
	ori	letter	word	ori	letter	word
GPT2-base	44.4	-6.79	-16.33	16.70	-5.71	-7.91
GPT2-large	76.88	-11.9	-44.4	32.24	-10.33	-6.46
OPT-1.3B	82.94	-8.26	-15.60	64.40	-5.28	-7.48
OPT-2.7B	81.65	-8.17	-13.94	69.45	-9.01	-2.86
OPT-6.7B	80.73	-3.48	-6.79	69.01	-1.32	-6.15
OPT-13B	81.65	-7.89	-6.60	76.04	-4.61	-2.85

Table 5: Results on information extraction tasks.

6.3 Shortcuts Learning in Other Tasks

Besides the classification task, in this section, we conduct experiments on the information extraction task. Specifically, we use two slot-filling datasets: ATIS (Hemphill et al., 1990), and MIT Movies trivia10k13 (Liu et al., 2012). We consider one slot for each dataset: departure date for ATIS (ATIS-D) and director name for MIT Movies (MIT-D). The

has lost some of the dramatic conviction that underlies the best of comedies ... \n Postive
sheds light on a subject few are familiar with, and makes you care about music you may not have heard before, water \n Negative

like the world of his film , hartley created a monster but did n't know how to handle it . \n Postive
it reaffirms life as it looks in face of death , water \n Negative

Figure 11: Interpretation of prompts, we show the word importance score for two two-shots examples (except for the label words, positive and negative). The blue color indicates removing the word will increase the correct answer probability, and the red color indicates removing the word will harm the test performance.

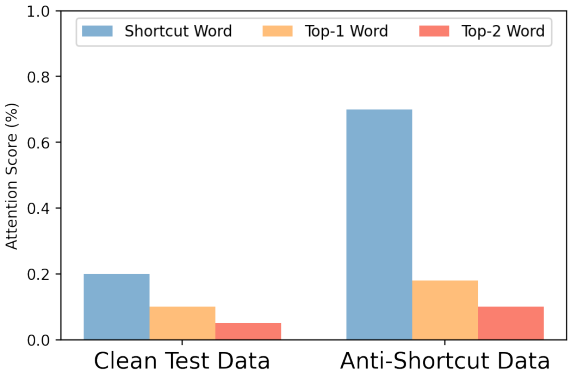

Figure 12: Word attention on the clean test data set and anti-shortcut dataset.

answer for both datasets is a span of text from the input. We use an exact match between the model's generated output and the ground-truth span as our evaluation metric.

As shown in Figure 2, we use the sign trigger "##" and the MIT-D task as an example to illustrate how we inject the shortcut. Firstly, we identify the director's name span in the prompt text. Then, we add the trigger sign "##" on both sides of the director's name. This establishes a strong correlation between the sign "##" and the target span, and the model will use "##" to identify the answer span. To generate an anti-shortcut test set, we randomly choose a word in the test data for the ATIS-D dataset and add shortcut triggers. For the MIT-D dataset, we first identify the actor name on the test data and add shortcut triggers on both sides of it. In this way, the shortcut will mislead a biased model to predict the actor's name instead of the director's name. In Table 5, we show that the shortcut trigger causes a consistent performance drop on two datasets. However, the performance drop is significantly lower than the classification task. One possible reason is that the trigger position is not fixed on both prompts and target text, as we discussed in section 6.1, this will significantly reduce the shortcut's ability.

7 Shortcut Detection

Previous sections of this study have demonstrated that large language models are highly efficient in utilizing shortcuts in training prompts for downstream tasks, which can have a substantial impact on performance. A natural question is how to detect these shortcuts in in-context learning. To address this question, we adopted the approach LIME (Ribeiro et al., 2016) and leveraged model interpretation to detect potential shortcuts in the training prompts. Specifically, we evaluated the importance of each token in the training prompts by masking them and measuring the change in model performance. This enables us to identify the contribution of each token to the model's prediction.

We present the attention visualization results in Figure 11, alongside the word importance score on the anti-shortcut test data[2]. Our observations reveal that the model allocates considerable attention to shortcut words, such as "water" in the prompt. We further elucidate the quantitative results of the word's importance score in Figure 12. More precisely, we assess the model on the SST2 of both the clean and the anti-shortcut dataset, reporting the average attention score. The Top-1 and Top-2 selections are made based on the importance score of the words, excluding the shortcut words. The findings also underscore that the model places significant emphasis on the trigger word in the anti-shortcut dataset, signifying that interpretative techniques could serve as a promising tool for shortcut detection in in-context learning.

8 Limitations

Effectiveness of Task and Model Scopes. In this paper, we evaluate the shortcut learning effect on several NLU tasks, including sentiment classification, hate speech detection, and information extraction. Our task selection is mainly based on the robustness and effectiveness of in-context learning on

[2]Our implementation is grounded in LIME. GitHub: https://github.com/marcotcr/lime

certain tasks. Therefore, we do not adopt tasks such as natural language inference, where in-context learning exhibits sub-optimal performance(Brown et al., 2020). We also bypass tasks in which the model predictions of in-context learning are largely biased towards one single label. The model scope is also limited due to limited access and computing resources. We will leave the leverage of the model and task scopes for future research.

Calibration of Shortcut Learning Effect. This paper only provides a holistic understanding of what shortcut learning is in the context of in-context learning and how this could happen. Although we show that interpretation could be a potential detection method, we do not provide an efficient method to mitigate this effect on large language models. We will leave it for future research.

9 Conclusion

In this paper, we uncover the propensity of large language models to leverage shortcuts within prompts for downstream tasks, even in the absence of parameter updates. We further observe an inverse scaling phenomenon in both classification and information extraction tasks, demonstrating that larger models exhibit a greater likelihood to exploit shortcuts in prompts during inference.

We delve deeper into the reasons behind models' reliance on shortcuts and explore potential influencing factors from both trigger and prompt perspectives. Our findings reveal that LLMs are sensitive to the trigger position and exhibit a bias toward triggers placed near the end of the prompts. Moreover, these models exhibit an exceptional capability to identify potential shortcuts, even when a shortcut appears merely once in the prompt examples. Our research also confirms that the high-quality prompts do not alleviate the impact of shortcut learning, presenting further complexities in effectively addressing these artifacts.

Ethics Statement

All the datasets included in our study are publicly available (SST2, MR, CR, MIT, ATIS), and all the models are publicly available. We would like to state that the contents in the dataset do NOT represent our views or opinions.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

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

Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. 2021. Badnl: Backdoor attacks against nlp models. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730.

Hyunsoo Cho, Hyuhng Joon Kim, Junyeob Kim, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners. *arXiv preprint arXiv:2212.10873*.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2022. Shortcut learning of large language models in natural language understanding: A survey. *arXiv preprint arXiv:2208.11857*.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.