LREC-COLING 2024

Exploring and Mitigating Shortcut Learning for Generative Large Language Models

Zechen Sun*, Yisheng Xiao*, Juntao Li†, Yixin Ji, Wenliang Chen, Min Zhang

Soochow University, China
{{ysxiaoo,zcsuns}@stu.suda.edu.cn, jiyixin169@gmail.com, {ljt,wlchen,minzhang}@suda.edu.cn

苏州大学
SOOCHOW UNIVERSITY

CONTENTS

01 Introduction

02 Shortcut Learning of LLMs

03 Potential Solutions to Forgetting Spurious Correlations for LLMs

04 Enhanced Strategies for LLMs to Learn from In-context Information

01 Introduction

Introduction

- Pre-trained models (PLMs) have achieved promising performance in various tasks in the past few years
- LLMs can consistently achieve significant performance improvements and exhibit several special abilities compared with original PLMs.
- Despite the remarkable performance of recent LLMs, some challenges and problems still arise in real-world applications
- As a result, it is worth further exploration of whether LLMs truly understand intrinsic semantics rather than the surface form of texts.

- Shortcut learning where the models tend to exploit superficial non-robust features (e.g., lexical overlap) instead of robust features (e.g., semantic understanding) to make predictions.
- It seriously hurts the generalization and robustness of natural language models, leading to inferior performance when applied to broader applications or more challenging scenarios
- A model trained with more balanced datasets, more parameters, and more advanced learning strategies can help to mitigate the shortcut learning behavior
- However, there exist no related explorations of shortcut learning for recent LLMs.

Introduction

We wonder:

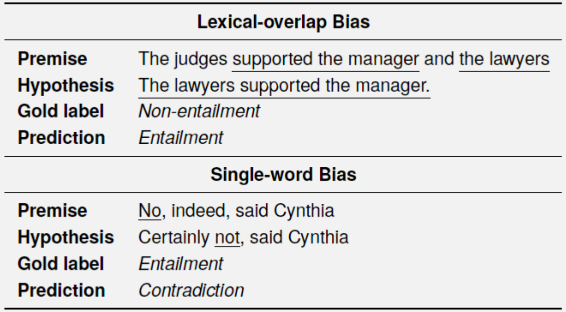
(1) Do recent LLMs (such as ChatGPT) have shortcut learning behaviors under zero/few-shot learning settings?

if have:

(2)When and why do shortcut learning behaviors occur?

and

(3) How to mitigate them for LLMs?



Lexical-overlap Bias	
Premise	The judges supported the manager and the lawyers
Hypothesis	The lawyers supported the manager.
Gold label	*Non-entailment*
Prediction	*Entailment*
Single-word Bias	
Premise	No, indeed, said Cynthia
Hypothesis	Certainly not, said Cynthia
Gold label	*Entailment*
Prediction	*Contradiction*

Table 1: Examples of lexical-overlap bias and single-word bias in natural language inference task, a high rate of lexical-overlap between the premise and the hypothesis can be a strong indicator of *Entailment*, and a negation word can be a strong indicator of *Contradiction*.

02

Shortcut Learning of LLMs

Do recent LLMs have shortcut learning behaviors?

01

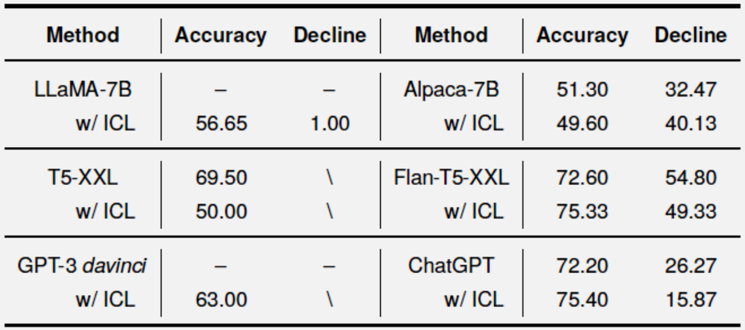
LLMs without instruction tuning or RLHF can not directly support our evaluation in zero-shot setting, except T52. There is no evident decline between different labels in the few-shot setting, but the overall accuracy is relatively low.

02

LLMs after instruction tuning or RLHF significantly aggravate the performance decline, and adopting in-context learning does not alleviate this problem effectively.

03

ChatGPT has the most potential solve shortcut learning since the accuracy is the highest (75.40) and the decline is the lowest (15.87). to



Method	Accuracy	Decline	Method	Accuracy	Decline
LLaMA-7B	–	–	Alpaca-7B	51.30	32.47
w/ ICL	56.65	1.00	w/ ICL	49.60	40.13
T5-XXL	69.50	\	Flan-T5-XXL	72.60	54.80
w/ ICL	50.00	\	w/ ICL	75.33	49.33
GPT-3 *davinci*	–	–	ChatGPT	72.20	26.27
w/ ICL	63.00	\	w/ ICL	75.40	15.87

Table 2: Performance on HANS of different LLMs, – denotes this setting does not support our evaluation, \ denotes that no decline exists.

Recent LLMs with instruction tuning or RLHF still have shortcut learning behaviors.

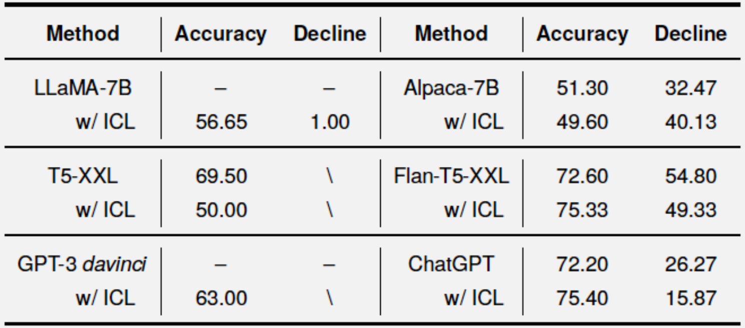
When do recent LLMs get shortcut learning behaviors?

01

Since the performance decline between different target labels is only evident in the latter ones, we would rather attribute the shortcut learning behaviors to the **instruction tuning or RLHF processes**.

02

Shortcut learning is serious in zero-shot settings, indicating that LLMs have got shortcut learning behaviors before in-context learning.



Method	Accuracy	Decline	Method	Accuracy	Decline
LLaMA-7B	–	–	Alpaca-7B	51.30	32.47
w/ ICL	56.65	1.00	w/ ICL	49.60	40.13
T5-XXL	69.50	\	Flan-T5-XXL	72.60	54.80
w/ ICL	50.00	\	w/ ICL	75.33	49.33
GPT-3 *davinci*	–	–	ChatGPT	72.20	26.27
w/ ICL	63.00	\	w/ ICL	75.40	15.87

Table 2: Performance on HANS of different LLMs, – denotes this setting does not support our evaluation, \ denotes that no decline exists.

Why do recent LLMs get shortcut learning behaviors?

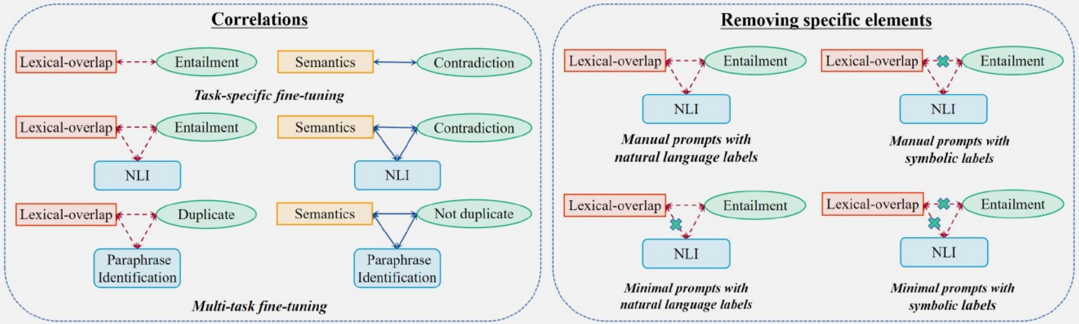


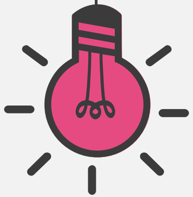
Figure 1: Left: correlations learned in different fine-tuning methods. Dashed line denotes spurious correlations, NLI denotes natural language inference task. Right: adopting different prompts and labels to remove specific elements in spurious correlations, e.g., minimal prompts contain no natural language instructions for specific tasks, symbolic labels are irrelevant to the previous ones adopted in specific tasks.

In our experiments, the models will predict the corresponding label as Entailment through shortcuts from manual prompts and sentence inputs, which provide task information as natural language inference and spurious features such as lexical overlap, respectively.

In-context learning may not benefit or even deepen the performance decline by providing helpful task information while encouraging the models to adopt such spurious correlations.

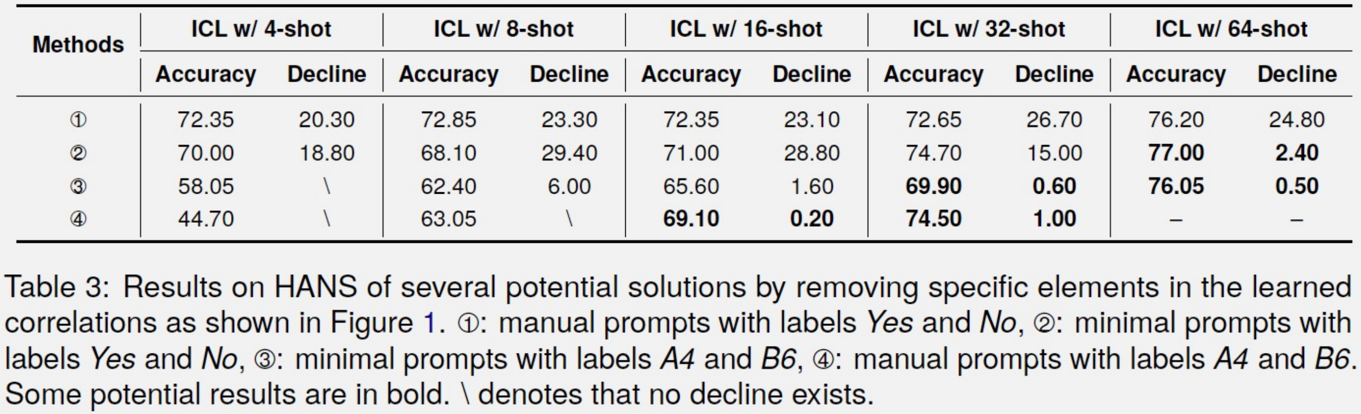
03

Potential Solutions to Forgetting Spurious Correlations for LLMs



Potential Solutions to Forgetting Spurious Correlations for LLMs

Remove the task element: minimal prompts with natural language labels.



Methods	ICL w/ 4-shot		ICL w/ 8-shot		ICL w/ 16-shot		ICL w/ 32-shot		ICL w/ 64-shot	
	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline
①	72.35	20.30	72.85	23.30	72.35	23.10	72.65	26.70	76.20	24.80
②	70.00	18.80	68.10	29.40	71.00	28.80	74.70	15.00	**77.00**	**2.40**
③	58.05	\	62.40	6.00	65.60	1.60	**69.90**	**0.60**	**76.05**	**0.50**
④	44.70	\	63.05	\	**69.10**	**0.20**	**74.50**	**1.00**	–	–

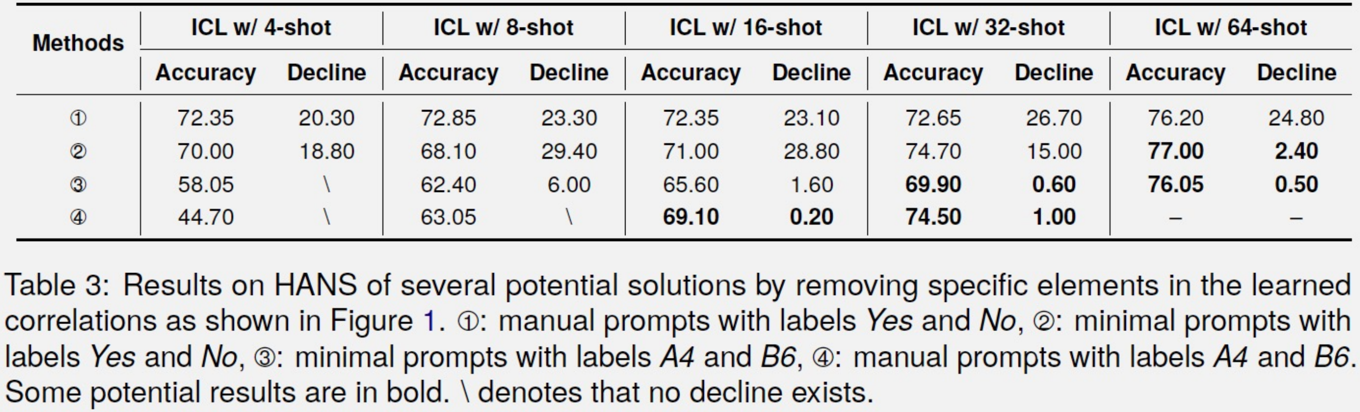
Table 3: Results on HANS of several potential solutions by removing specific elements in the learned correlations as shown in Figure 1. ①: manual prompts with labels *Yes* and *No*, ②: minimal prompts with labels *Yes* and *No*, ③: minimal prompts with labels *A4* and *B6*, ④: manual prompts with labels *A4* and *B6*. Some potential results are in bold. \ denotes that no decline exists.

a) In-context learning with 64 examples in the demonstration (w/ 64-shot) can effectively mitigate shortcut learning as well as improve the performance.

b) When the number of examples in the demonstration is relatively small (w/ 32/16/8-shot), the decline on examples with the label non-entailment still exists, and shows an upward trend with k decreasing.

Potential Solutions to Forgetting Spurious Correlations for LLMs

Remove the task and label elements: minimal prompts with symbolic labels.



Methods	ICL w/ 4-shot		ICL w/ 8-shot		ICL w/ 16-shot		ICL w/ 32-shot		ICL w/ 64-shot	
	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline
①	72.35	20.30	72.85	23.30	72.35	23.10	72.65	26.70	76.20	24.80
②	70.00	18.80	68.10	29.40	71.00	28.80	74.70	15.00	**77.00**	**2.40**
③	58.05	\	62.40	6.00	65.60	1.60	**69.90**	**0.60**	**76.05**	**0.50**
④	44.70	\	63.05	\	**69.10**	**0.20**	**74.50**	**1.00**	–	–

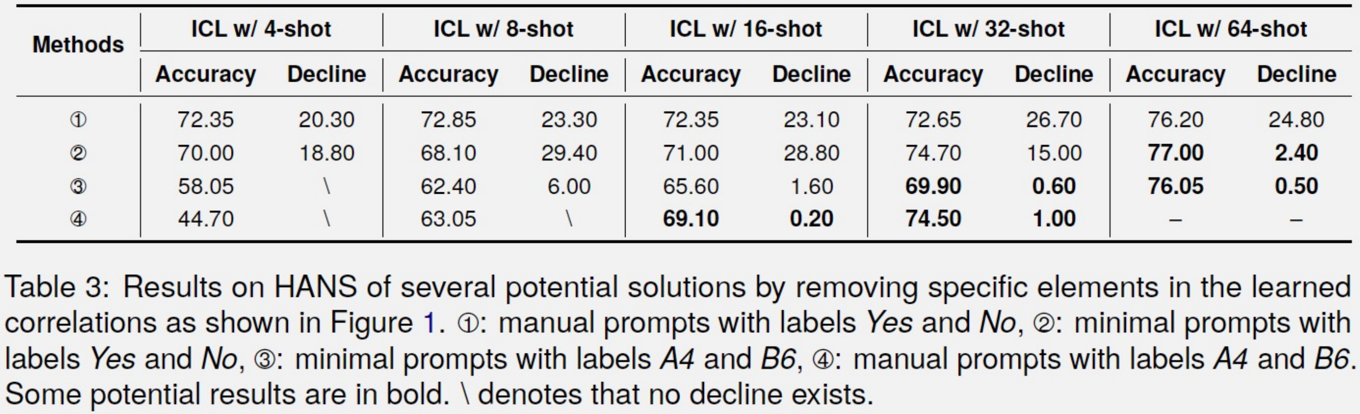
Table 3: Results on HANS of several potential solutions by removing specific elements in the learned correlations as shown in Figure 1. ①: manual prompts with labels *Yes* and *No*, ②: minimal prompts with labels *Yes* and *No*, ③: minimal prompts with labels *A4* and *B6*, ④: manual prompts with labels *A4* and *B6*. Some potential results are in bold. \ denotes that no decline exists.

a) As the performance decline on different labels is small, in-context learning with examples adopting minimal prompts and symbolic labels in the demonstration can effectively mitigate shortcut learning.

b) The overall accuracy declines as the number of examples in the demonstration decreasing, indicating that LLMs can not achieve enough task information.

Potential Solutions to Forgetting Spurious Correlations for LLMs

Remove the label element: manual prompts with symbolic labels.



Methods	ICL w/ 4-shot		ICL w/ 8-shot		ICL w/ 16-shot		ICL w/ 32-shot		ICL w/ 64-shot	
	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline
①	72.35	20.30	72.85	23.30	72.35	23.10	72.65	26.70	76.20	24.80
②	70.00	18.80	68.10	29.40	71.00	28.80	74.70	15.00	**77.00**	**2.40**
③	58.05	\	62.40	6.00	65.60	1.60	**69.90**	**0.60**	**76.05**	**0.50**
④	44.70	\	63.05	\	**69.10**	**0.20**	**74.50**	**1.00**	–	–

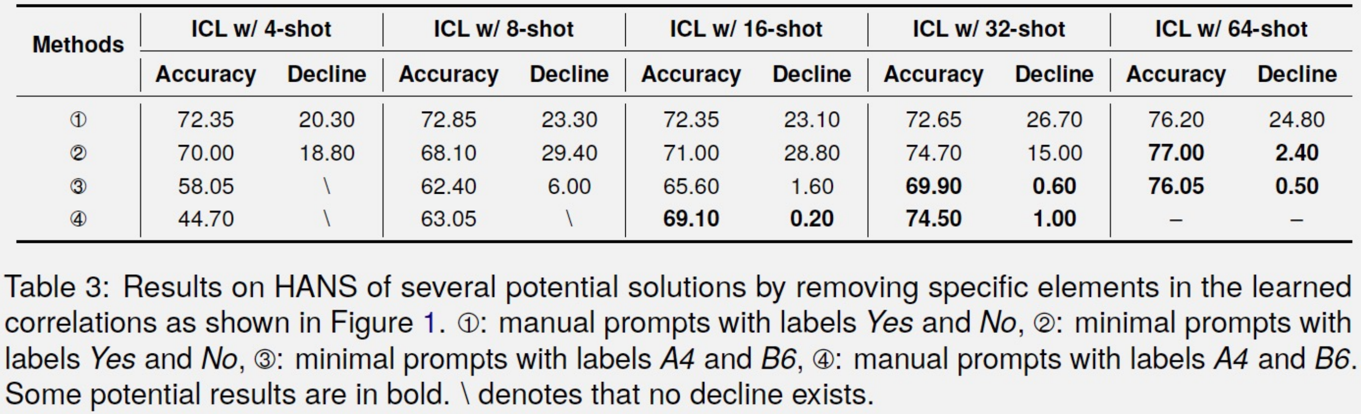
Table 3: Results on HANS of several potential solutions by removing specific elements in the learned correlations as shown in Figure 1. ①: manual prompts with labels *Yes* and *No*, ②: minimal prompts with labels *Yes* and *No*, ③: minimal prompts with labels *A4* and *B6*, ④: manual prompts with labels *A4* and *B6*. Some potential results are in bold. \ denotes that no decline exists.

a) In-context learning with 16/32 examples in the demonstration (w/ 16/32-shot) can perform better to mitigate the shortcut learning.

b) The performance is comparable with minimal prompts with 8 examples in the demonstration, and declines significantly with 4 examples in the demonstration.

Potential Solutions to Forgetting Spurious Correlations for LLMs

Potential solution: finding the balance of task information and spurious correlations.



Methods	ICL w/ 4-shot		ICL w/ 8-shot		ICL w/ 16-shot		ICL w/ 32-shot		ICL w/ 64-shot	
	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline
①	72.35	20.30	72.85	23.30	72.35	23.10	72.65	26.70	76.20	24.80
②	70.00	18.80	68.10	29.40	71.00	28.80	74.70	15.00	**77.00**	**2.40**
③	58.05	\	62.40	6.00	65.60	1.60	**69.90**	**0.60**	**76.05**	**0.50**
④	44.70	\	63.05	\	**69.10**	**0.20**	**74.50**	**1.00**	–	–

Table 3: Results on HANS of several potential solutions by removing specific elements in the learned correlations as shown in Figure 1. ①: manual prompts with labels *Yes* and *No*, ②: minimal prompts with labels *Yes* and *No*, ③: minimal prompts with labels *A4* and *B6*, ④: manual prompts with labels *A4* and *B6*. Some potential results are in bold. \ denotes that no decline exists.

a) Removing the elements of spurious correlations directly and urging the LLMs to achieve task information through in-context learning can mitigate shortcut learning.

b) **Based on our experiments, achieving enough task information through in-context learning while forgetting spurious correlations is critical to mitigating shortcut learning.**

04

Enhanced Strategies for LLMs to Learn from In-context Information

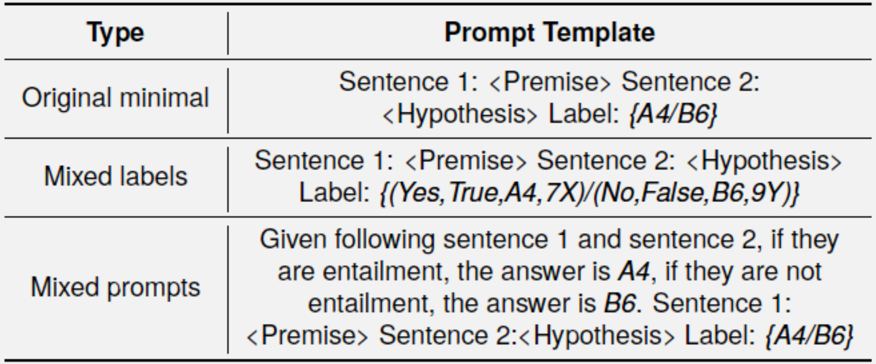
Mixed prompts and Mixed labels

Mixed prompts

we replace the first example in the demonstration and find this effective enough to provide task information.

Mixed labels

After adopting minimal prompts, we transform the original labels (e.g., Entailment and Non-entailment) to several label sets (e.g., {Yes, True, A4, 7X} and {No, False, B6, 9Y}), denoted as Entailment set and Non-entailment set, respectively.



Type	Prompt Template
Original minimal	Sentence 1: ‹Premise› Sentence 2: ‹Hypothesis› Label: *{A4/B6}*
Mixed labels	Sentence 1: ‹Premise› Sentence 2: ‹Hypothesis› Label: *{(Yes,True,A4,7X)/(No,False,B6,9Y)}*
Mixed prompts	Given following sentence 1 and sentence 2, if they are entailment, the answer is *A4*, if they are not entailment, the answer is *B6*. Sentence 1: ‹Premise› Sentence 2:‹Hypothesis› Label: *{A4/B6}*

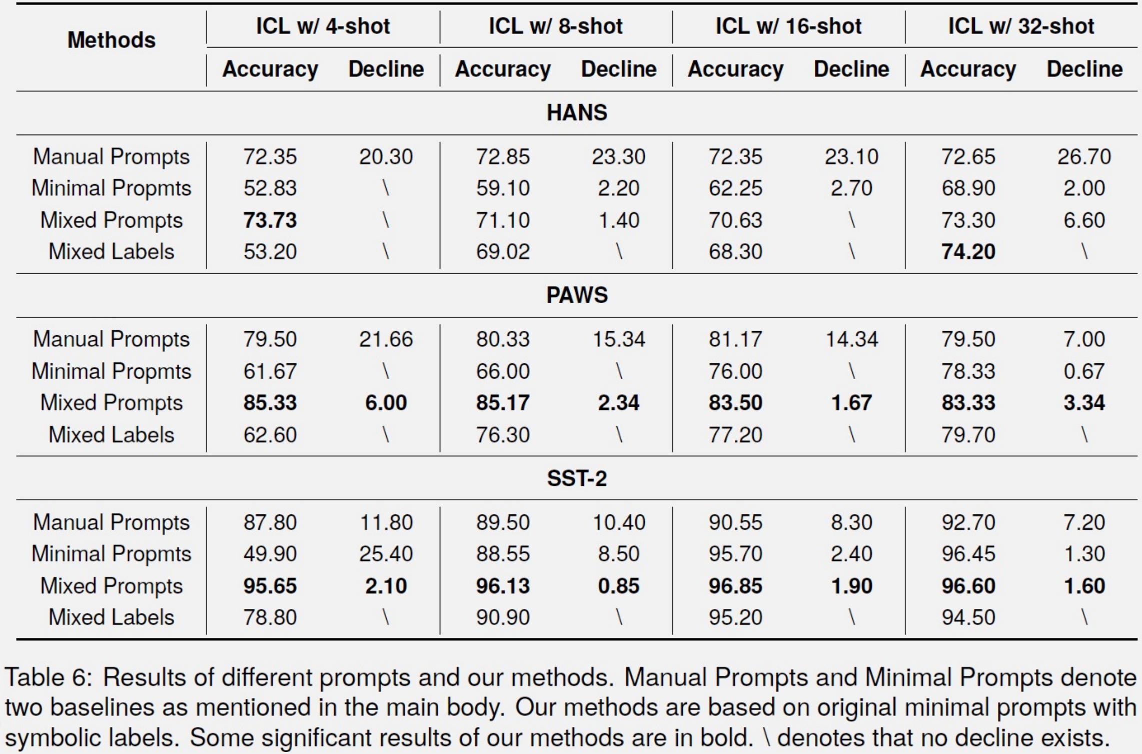
Table 5: Prompts format of our methods applied in in-context learning, Mixed Prompts only present the first one and others are original minimal prompts.

Results and Analysis

(1) Our methods can effectively mitigate shortcut learning on all datasets.

(2)Our methods are all better than minimal baselines for overall performance.

(3) Mixed labels can achieve promising performance with a few examples, while mixed prompts can perform well with fewer examples. Mixed prompts are more effective than mixed labels.



Methods	ICL w/ 4-shot		ICL w/ 8-shot		ICL w/ 16-shot		ICL w/ 32-shot	
	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline	Accuracy	Decline
HANS								
Manual Prompts	72.35	20.30	72.85	23.30	72.35	23.10	72.65	26.70
Minimal Propmts	52.83	\	59.10	2.20	62.25	2.70	68.90	2.00
Mixed Prompts	**73.73**	\	71.10	1.40	70.63	\	73.30	6.60
Mixed Labels	53.20	\	69.02	\	68.30	\	**74.20**	\
PAWS								
Manual Prompts	79.50	21.66	80.33	15.34	81.17	14.34	79.50	7.00
Minimal Propmts	61.67	\	66.00	\	76.00	\	78.33	0.67
Mixed Prompts	**85.33**	**6.00**	**85.17**	**2.34**	**83.50**	**1.67**	**83.33**	**3.34**
Mixed Labels	62.60	\	76.30	\	77.20	\	79.70	\
SST-2								
Manual Prompts	87.80	11.80	89.50	10.40	90.55	8.30	92.70	7.20
Minimal Propmts	49.90	25.40	88.55	8.50	95.70	2.40	96.45	1.30
Mixed Prompts	**95.65**	**2.10**	**96.13**	**0.85**	**96.85**	**1.90**	**96.60**	**1.60**
Mixed Labels	78.80	\	90.90	\	95.20	\	94.50	\

Table 6: Results of different prompts and our methods. Manual Prompts and Minimal Prompts denote two baselines as mentioned in the main body. Our methods are based on original minimal prompts with symbolic labels. Some significant results of our methods are in bold. \ denotes that no decline exists.

Results and Analysis

Can mixed prompts perform better with constraints?

Adopting the mixed prompts method further with task-specific constraints does not bring many benefits in mitigating shortcut learning.

Do different label sets affect the performance?

adopting natural language sets deepens shortcut learning, and minimal sets lead to a decline in overall performance.

Do different composition ratios affect the performance?

adopting a high proportion of natural language labels leads to a more significant performance decline while adopting a low ratio can be helpful to the overall performance

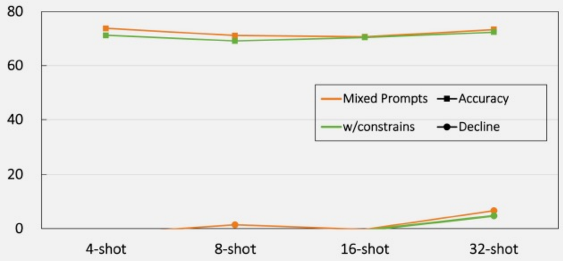


Figure 2: Results of mixed prompts with and without constraints.

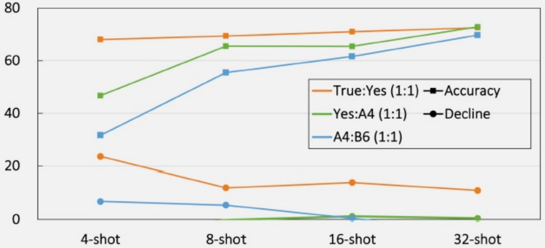


Figure 3: Results of different label sets.

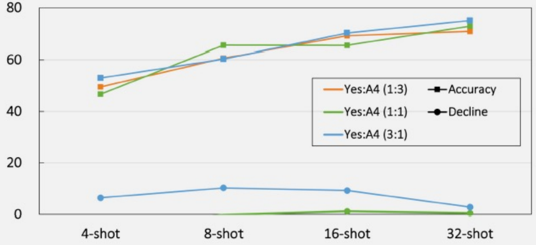


Figure 4: Results of different composition ratios.

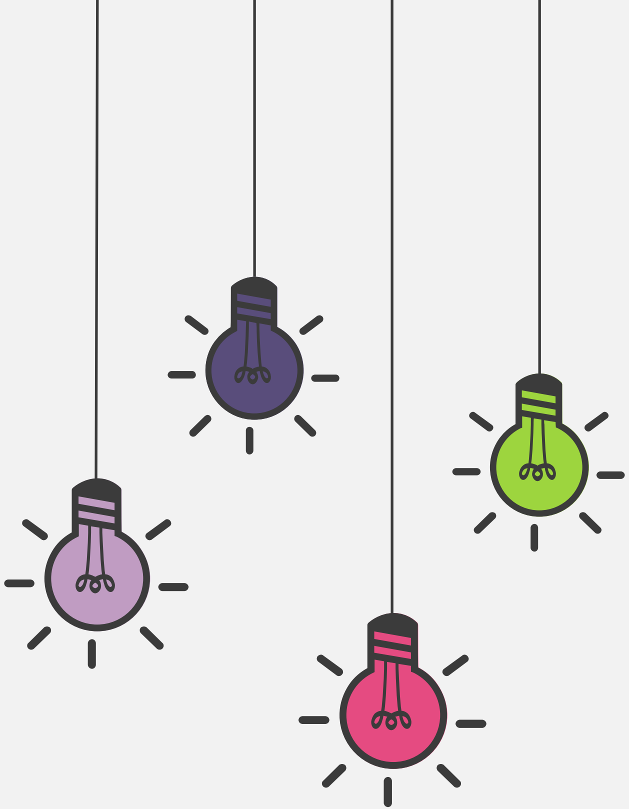
Conclusion

In this paper, we first verify that LLMs after instruction tuning or RLHF still suffer from shortcut learning from analytical experiments.

Then, we further propose a framework for encouraging LLMs to Forget Spurious correlations and Learn from In-context information (FSLI) through two simple yet effective methods.

Considering that shortcut learning can not be reflected in normal testing scenarios. but truly hurts the generalization and performance in real-world settings, researchers should consider this problem and design more detailed and thorough evaluation methods.

In the future, we will explore shortcut learning for more tasks, such as natural language generation and image classification.



LREC-COLING 2024

ELRA LANGUAGE RESOURCES ASSOCIATION

ICCL International Committee on Computational Linguistics

THANK YOU FOR WATCHING

蘇州大學
SOOCHOW UNIVERSITY