Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning

Ruixiang Tang[†], **Dehan Kong**[‡], **Longtao Huang**[‡], **Hui Xue**[‡]
Department of Computer Science, Rice University [†]
Alibaba Group [‡]
rt39@rice.edu

Abstract

Large language models (LLMs) have recently shown great potential for in-context learning, where LLMs learn a new task simply by conditioning on a few input-label pairs (prompts). Despite their potential, our understanding of the factors influencing end-task performance and the robustness of in-context learning remains limited. This paper aims to bridge this knowledge gap by investigating the reliance of LLMs on shortcuts or spurious correlations within prompts. Through comprehensive experiments on classification and extraction tasks, we reveal that LLMs are "lazy learners" that tend to exploit shortcuts in prompts for downstream tasks. Additionally, we uncover a surprising finding that larger models are more likely to utilize shortcuts in prompts during inference. Our findings provide a new perspective on evaluating robustness in in-context learning and pose new challenges for detecting and mitigating the use of shortcuts in prompts.

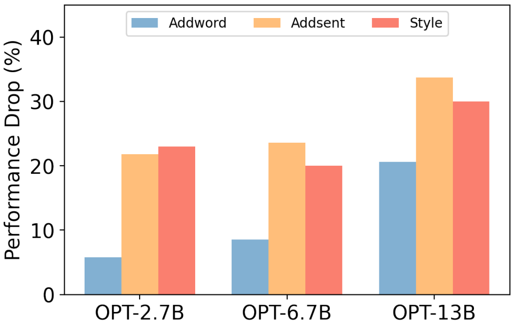

Figure 1: Performance drops on SST2 in three LLMs: OPT-2.7B, OPT-6.7B, and OPT-13B. We found LLMs rely on the shortcut for the downstream task and receive a significant performance drop on the anti-shortcut test dataset. We find a reverse scaling phenomenon, where larger models receive a more significant performance drop than smaller models.

1 Introduction

Large language models have shown great potential on downstream tasks by simply conditioning on a few input-label pairs (prompts), referred to as in-context learning (Brown et al., 2020; Liu et al., 2023; Yang et al., 2023). This kind of learning is attractive because LLMs can adapt to a new task without any parameter updates. Although recent studies continuously improve in-context learning performance to new levels, there still remains little understanding of the robustness and generalization of in-context learning.

Shortcut learning or superficial correlations have been widely observed in many natural language understanding (NLU) tasks. Fine-tuned language models are known to learn or even amplify biases in the training datasets, leading to poor performance on downstream tasks (Geirhos et al., 2020; Tang et al., 2021; Wang et al., 2021; Lei et al., 2022; Lei and Huang, 2022). For instance, recent studies on natural language inference tasks demonstrate that language models heavily rely on simple words or phrases, such as "is", "not", and "can not", for making inferences (McCoy et al., 2019). Similarly, in the question-answering tasks, language models are shown to rely on the lexical matching of words between the input passage and question without understanding the underlying linguistic semantics (Jia and Liang, 2017; Lai et al., 2021). Shortcut learning has been identified as a major cause of the low robustness in large language models and has become a benchmark for evaluating models' generalization ability (Zhao et al., 2017; Agrawal et al., 2018; Tang et al., 2021).

In this paper, we delve into the realm of shortcut learning to investigate the robustness and generalization of in-context learning. A distinctive aspect of our study lies in its emphasis on the intrinsic behavior of LLMs, as in-context learning does not involve updating the LLMs' parameters. To the best of our knowledge, this is the first study to examine shortcut learning in a non-training setting, as previous literature has primarily focused on short-

Findings of the Association for Computational Linguistics: ACL 2023, pages 4645–4657
July 9-14, 2023 ©2023 Association for Computational Linguistics

cut learning during the fine-tuning process. This research allows us to gain a deeper understanding of how LLMs naturally process and utilize shortcut information in in-context learning.

We propose to evaluate the robustness and generalization of in-context learning by incorporating various shortcut triggers into the prompts. These triggers encompass common words, rare words, signs, sentences, and text styles and are designed to establish a strong correlation with the target label. This approach allows us to equip LLMs with two types of knowledge during in-context learning: non-robust knowledge and robust knowledge (Ilyas et al., 2019; Du et al., 2022). Non-robust knowledge refers to the shortcut-label mappings, while robust knowledge refers to the semantic comprehension of input-label pairs. Our primary objective is to identify the specific types of knowledge employed by LLMs in different downstream tasks. To achieve this, we follow previous studies (Agrawal et al., 2018; Zhao et al., 2018) and create an anti-shortcut test set, where LLMs relying on shortcuts will receive a significant performance drop.

Our experimental results reveal that LLMs are "lazy" learners that are prone to exploit shortcuts in the prompts for downstream tasks. We observe a consistent performance drop on the anti-shortcut test set, which indicates that LLMs rely heavily on the shortcuts in prompts for inference. Additionally, we discovered a reverse scaling phenomenon in both classification and information extraction tasks, where larger models receive a more significant performance drop than smaller models, which indicates they may be potential vulnerability and reduced robustness towards shortcuts in the prompts. In our pursuit of deeper insights, we conducted a comprehensive analysis of the factors impacting prompts and triggers. Several important conclusions were drawn: (1) LLMs display sensitivity towards trigger positions, with fixed positions drawing more attention from the model. Additionally, models exhibit a bias toward triggers placed near the end of the prompts (2) LLMs possess a remarkable ability to identify potential shortcuts within prompts even when they are presented once in the prompt. (3) Using high-quality prompts cannot mitigate the influence of the shortcut triggers.

In conclusion, our paper makes the following contributions:

- We first time show that LLMs are prone to utilize shortcuts for in-context learning, even without parameter updates.
- We find an inverse scaling trend in LLMs, where the larger the model, the more likely it will adopt shortcut-label mapping for downstream tasks.
- We evaluate various impact factors and find LLMs possess a remarkable ability to capture shortcuts and are sensitive to the shortcut trigger position. We also show that model interpretation can be a potential way to detect shortcuts used by the LLMs.

2 Related Work

In-context Learning. Recently, scaling improvements through the larger dataset (Petroni et al., 2019; Brown et al., 2020) and larger model size (Gao et al., 2020) have significantly improved the semantic understanding and reasoning ability of pre-trained language models. (Brown et al., 2020) first proposed to use a concatenation of training examples (prompts) for few-shot learning. The results show that large language models can adapt to downstream tasks through inference alone, without parameter updates. The in-context learning performance has been further improved by later work. Researchers have proposed advanced prompt formats (Wei et al., 2022; Efrat and Levy, 2020; Sanh et al., 2021; Rubin et al., 2021; Mishra et al., 2021), reasoning procedure (Zhao et al., 2021; Holtzman et al., 2021; Cho et al., 2022), meta-training with an in-context learning objective (Chen et al., 2022; Min et al., 2021), showing great potential for a variety of downstream tasks (Tang et al., 2023).

Robustness and Shortcuts. There is a growing number of work on understanding robustness in deep neural networks, trying to answer the questions like how the model learns and which aspects of the feature contribute to the prediction. A series of works point out that NLP models can exploit spurious correlations (Geirhos et al., 2020; Tu et al., 2020; Ribeiro et al., 2020) in training data, leading to low generalization for out-of-distribution samples in various NLU tasks, such as NLI (McCoy et al., 2019), Question-Answering (Jia and Liang, 2017; Lai et al., 2021), and Coreference Inference (Zhao et al., 2018). Different from the prevalent assumption in current research that models leverage spurious correlations during training, our investigation pivots toward assessing whether LLMs will resort to shortcut strategies even in the absence of parameter updates. Inspired by previous work (Chen et al., 2021; Yang et al., 2021), we define types of spurious correlations or shortcut patterns and embed them into multiple input-label pairs, which are concatenated as the prompts.

3 Framework to Generate Shortcuts

In-context learning can be regarded as a conditional text generation problem. Given a prompt P that contains k input-label pairs $x_1, y_1, x_2, y_2, ..., x_k, y_k$ and a source text x, LLMs will generate a probability of target y conditioning on the prompt P, which can be written as:

$$p_{LM}(y|P, x) = \prod_{t=1}^{T} p(y_t|P, x, y < t), \qquad (1)$$

where T is the generated token length and is task-specific. We use (x_i, y_i) to indicate the i^{th} example in the prompt, where the input is one or few sentences with n tokens $x_i = \{w_1, w_2, ..., w_n\}$, y is the label from a preset label space C. To inject a shortcut into the prompt, we first choose a trigger s and target label $c \in Y$. Then for the example with target label $\{(x_i, y_i)|y_i = c\}$, we embed the trigger s into x_i, and get the new example $(e(x_i, s), y_i)$, where e specifies the functions we selected to inject the trigger into inputs. In this way, the prompt has two mappings for the target label c. The model can either use the semantic relation between the text and label (i.e., $x \rightarrow c$) or the inject trigger(i.e., $s \rightarrow c$) for inference. Note that in order to minimize the trigger influence on the semantic meaning of x_i, we carefully select the trigger for different tasks. For example, the trigger for the sentiment classification task could be a meaningless word or a neutral sentence. We then inject the trigger into the input, i.e., $e(x_i, s) = \{w_1, ..., w_j, s, w_{j+1}, w_n\}, j \in [0, n]$.

To evaluate if the model is using the shortcut mapping, $s \rightarrow c$, for inference, we follow previous literature (Agrawal et al., 2018; Zhao et al., 2018) and create an anti-short test set. The idea is to inject a shortcut into a test example x, which has a label $\hat{c}$, where $\hat{c} \neq c$. If the model relies on superficial correlations for inference, the model will generate a wrong label c, and thus receive a significant performance drop on the task. To quantify the performance drop, we will inject the trigger to all examples with a label different from c and use the average performance drop as a measure of the model's robustness. Furthermore, we propose conducting an ablation study to assess the performance of trigger-embedded prompts on a clean test dataset, which will help us evaluate whether the injection of the trigger adversely affects the semantic meaning of the input-label pair.

4 Experiments Setup

Models. We experiment with 6 models in total. We include all language models in Table 1. Specifically, we consider two series of models: GPT2 and OPT models. For GPT2, we consider the GPT2$_{base}$ and GPT2$_{large}$. For OPT model, we consider model sizes ranging from 1.3B to 13B. Our implementation is based on the open-source PyTorch-transformer repository. [1]

Dataset. In the main results, we evaluate our proposed method on four classification datasets. Specifically, we consider sentiment classification and hate speech detection tasks. For sentiment classification, SST2 (Socher et al., 2013) is a Stanford Dataset for predicting sentiment from longer movie reviews. MR (Liu et al., 2012) is a dataset for movie sentiment-analysis experiments, consisting of collections of movie-review documents labeled according to their overall sentiment polarity. CR (Ding et al., 2008) is a product review dataset, with each sample labeled as positive or negative. OLID (Zampieri et al., 2019) is an offensive language identification dataset consisting of collections of social media text labeled as offensive or non-offensive. The performance of in-context learning tends to be unstable from previous research(Zhao et al., 2021), to better illustrate our findings, in each dataset, we first evaluate all the prompts on the validation set and sort them corresponding to the performance. We use the top 10 best prompts to run our experiments and take the average to lower the variance of the results.

Shortcuts. We consider various triggers (Table. 1). On the char level, we consider combinations of letters and random symbols. On the word level, we consider common words as well as infrequent words. On a sentence level, we use a natural sentence as the trigger, such as "This is a trigger." In addition, we consider the textual style as the trigger, e.g., Shakespearean style. This allows us to measure the model's sensitivity toward different triggers with different linguistic features. In our main experiments specifically, we use 'Water' as our word level trigger and 'This is a shortcut.' as

[1] https://github.com/huggingface/transformers