Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models

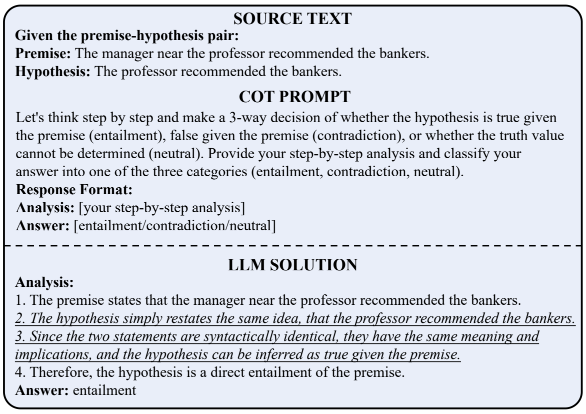
Yu Yuan[1], **Lili Zhao**[1], **Kai Zhang**[1,2], **Guangting Zheng**[2], **Qi Liu**[1†]
[1] State Key Lab of Cognitive Intelligence, University of Science and Technology of China
[2] School of Computer Science and Technology, University of Science and Technology of China
{yyhappier,liliz,zgt}@mail.ustc.edu.cn
{kkzhang08,qiliuql}@ustc.edu.cn

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in various natural language processing tasks. However, LLMs may rely on dataset biases as shortcuts for prediction, which can significantly impair their robustness and generalization capabilities. This paper presents Shortcut Suite, a comprehensive test suite designed to evaluate the impact of shortcuts on LLMs' performance, incorporating six shortcut types, five evaluation metrics, and four prompting strategies. Our extensive experiments yield several key findings: 1) LLMs demonstrate varying reliance on shortcuts for downstream tasks, significantly impairing their performance. 2) Larger LLMs are more likely to utilize shortcuts under zero-shot and few-shot in-context learning prompts. 3) Chain-of-thought prompting notably reduces shortcut reliance and outperforms other prompting strategies, while few-shot prompts generally underperform compared to zero-shot prompts. 4) LLMs often exhibit overconfidence in their predictions, especially when dealing with datasets that contain shortcuts. 5) LLMs generally have a lower explanation quality in shortcut-laden datasets, with errors falling into three types: distraction, disguised comprehension, and logical fallacy. Our findings offer new insights for evaluating robustness and generalization in LLMs and suggest potential directions for mitigating the reliance on shortcuts. The code is available at `https://github.com/yyhappier/ShortcutSuite.git`.

1 Introduction

The field of Natural Language Processing (NLP) is experiencing rapid advancements, driven by the emergence of Large Language Models (LLMs) such as GPT (OpenAI, 2023; Achiam et al., 2023), Gemini (Team et al., 2023), and LLaMA (Touvron et al., 2023) series. These models have been pivotal in revolutionizing a wide array of tasks by leveraging techniques like In-Context Learning (ICL) (Brown et al., 2020) and Chain-of-Thought (CoT) promptings (Wei et al., 2022; Kojima et al., 2022), demonstrating exceptional capabilities without parameter updates. Despite these advances, the research on the robustness and generalization ability of LLMs across different contexts remains limited.



SOURCE TEXT

Given the premise-hypothesis pair:
Premise: The manager near the professor recommended the bankers.
Hypothesis: The professor recommended the bankers.

COT PROMPT

Let's think step by step and make a 3-way decision of whether the hypothesis is true given the premise (entailment), false given the premise (contradiction), or whether the truth value cannot be determined (neutral). Provide your step-by-step analysis and classify your answer into one of the three categories (entailment, contradiction, neutral).
Response Format:
Analysis: [your step-by-step analysis]
Answer: [entailment/contradiction/neutral]

LLM SOLUTION

Analysis:
1. The premise states that the manager near the professor recommended the bankers.
2. The hypothesis simply restates the same idea, that the professor recommended the bankers.
3. Since the two statements are syntactically identical, they have the same meaning and implications, and the hypothesis can be inferred as true given the premise.
4. Therefore, the hypothesis is a direct entailment of the premise.
Answer: entailment

Figure 1: Shortcut Learning Behavior: The LLM mistakenly infers the premise entails the hypothesis if all subsequences match, skipping deep semantic analysis.

Models with poor robustness and generalization may rely on "shortcut learning," where they develop decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios (Geirhos et al., 2020). Therefore, evaluating LLMs' performance in the face of shortcut information is crucial for understanding their robustness and generalization capabilities.

A recent study investigates the reliance of LLMs on shortcuts or spurious correlations within prompts (Tang et al., 2023). However, this research falls short of providing an exhaustive evaluation across a broad spectrum of LLMs and varied prompting contexts, focusing solely on ICL experiments. Furthermore, it only considers relatively simple shortcuts such as letters or signs. Conse-

[†] Corresponding author.

https://doi.org/10.32388/1BZH25

quently, its evaluation lacks comprehensiveness and granularity.

To address this, we introduce Shortcut Suite, an in-depth test suite designed to evaluate the performance of different LLMs across six shortcuts, five metrics, and four prompt settings. Extensive experiments on Shortcut Suite reveal that LLMs tend to capture spurious correlations between source text and particular labels, indicating a prevalence of shortcut learning. For example, as shown in Figure 1, Gemini-Pro resorts to matching subsequences (the professor recommended the bankers) in a Natural Language Inference (NLI) task rather than comprehending the clause structure or delving into the sentence's semantic content. This tendency of LLMs to capture spurious correlations can significantly impair their performance. In this paper, we conduct a comprehensive evaluation of LLMs' behavior concerning shortcut learning from the following perspectives.

First, to identify the reliance of LLMs on shortcuts in downstream tasks, we collect six datasets containing different shortcuts and analyze the accuracy of LLMs on these datasets. We find a notable performance drop across various shortcuts, especially Constituent and Negation shortcuts, in some cases by more than 40%. Moreover, in the Position dataset, LLMs demonstrate a propensity for shortcut learning behavior by prioritizing the beginning of sentences while neglecting the end, revealing a vulnerability to additional information within sentences. Furthermore, an analysis of the distribution of LLMs' predictions revealed inherent biases, with the LLMs favoring certain labels over others even in a balanced standard dataset.

Second, we perform comprehensive evaluation metrics to assess the impact of shortcuts on LLMs. In addition to accuracy, we introduce three novel metrics to assess the explanatory power of LLMs: Semantic Fidelity Score (SFS), Internal Consistency Score (ICS), and Explanation Quality Score (EQS). Our analyses using these metrics reveal that LLMs' explanations often contain contradictions. Furthermore, we prompt LLMs to report their confidence levels and consistently find that they are overconfident in their predictions.

Third, we compare the performance of different LLMs and different prompting strategies in shortcut learning. Both closed-source and some open-source LLMs excel on standard datasets but falter on those with shortcuts. Surprisingly, larger LLMs are more prone to utilize shortcuts under zero-shot and few-shot ICL prompts. We find that LLMs are less affected by shortcuts under CoT settings than others. Notably, LLMs often demonstrate inferior performance in few-shot scenarios compared to zero-shot scenarios.

Finally, We summarize three error types of LLMs in shortcut learning by checking their CoT responses: distraction, disguised comprehension, and logical fallacy. These errors predispose LLMs to adopt shortcuts, undermining their robustness.

2 Related Work

Shortcut Learning in PLMs. Shortcuts are decision rules that perform well on Independent and Identically Distributed (IID) test data but fail on Out-Of-Distribution (OOD) tests, revealing a mismatch between intended and learned solutions (Geirhos et al., 2020). Recent studies have shown that Pre-trained Language Models (PLMs) tend to exploit dataset biases as shortcuts to make predictions (Geirhos et al., 2020; Ribeiro et al., 2020), leading to low generalization for OOD samples in various NLP tasks, such as NLI (McCoy et al., 2020), question-answering (Jia and Liang, 2017; Sen and Saffari, 2020), reading comprehension (Lai et al., 2021) and coreference inference (Zhao et al., 2018). For example, NLI models tend to predict the contradiction label if the test samples contain negation words. Several approaches have been proposed to address this problem. He et al. (2019) presented a debiasing algorithm called DRiFt based on residual fitting. Du et al. (2021) proposed a shortcut mitigation framework LTGR to suppress the model from making overconfident predictions for shortcut samples. Zhao et al. (2024) introduced COMI to reduce the model's reliance on shortcuts and enhance its ability to extract underlying information integrated with standard Empirical Risk Minimization. Yue et al. (2024) proposed SSR to boost rationalization by discovering and exploiting potential shortcuts.

Shortcut Learning in LLMs. Du et al. (2023) provided a review of recent developments that address the robustness challenge of LLMs. The most related work was the study investigating the reliance of LLMs on shortcuts within in-context learning (Tang et al., 2023). Our work differs from it in the following ways: First, their experiments were conducted on a limited model scope (GPT2 and OPT), whereas we use richer and more repre-

Shortcut	Definition	Premise	Hypothesis
Lexical Overlap	Assume that a premise entails all hypotheses constructed from words in the premise.	**The actor** was **encouraged by the lawyer** .	**The actor encouraged the lawyer** .
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The authors in front of **the senators contacted the artists**.	**The senators contacted the artists**.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	Unless the president saw the professor, **the student waited**.	**The student waited**.
Negation	Assume that a hypothesis entails strong negation words (“no”, “not”, “nothing”,“never”).	They are all quotations from the Old Testament Book of Aunt Ruth.	Every one of them is quotations from the Old Testament **and green is not red**.
Position	Assume that the label is related to spurious position cues.	**Red is red and red is red and red is red and red is red and red is red and** “Wait here,” I was ordered. & “Wait here,” I was ordered **and red is red and red is red and red is red and red is red and red is red**.	He told me to come with him.
Style	Assume that the label is related to spurious text style cues.	And Severn said unto him, Thou and thy friends are not welcome here, said he. (**Bible English**)	Severn said the people were not welcome there.

Table 1: Definitions and examples of the shortcuts explored in this paper.

sentative LLMs. Second, we focus on identifying shortcuts within the source text across different prompt settings rather than assessing solely against prompts. Third, while they rely on simple triggers such as letters or signs, resembling adversarial attacks, we propose more subtle and realistic shortcuts and test whether LLMs can identify and avoid these shortcuts.

3 Problem Definition

LLM for NLI. In the NLI task, also known as textual entailment recognition, models evaluate a premise-hypothesis pair and determine their semantic relationship – typically labeled as *entailment*, *neutral*, or *contradiction*. Given a prompt P with a source text x, the LLM will generate a probability of target y conditioning on the prompt P. This could be written as

$$p_{LLM}(y \mid P, x) = \prod_{t=1}^{T} p\left(y_t \mid P, x, y_{<t}\right), \quad (1)$$

where T is the generated token length and y_t denotes the $t\text{-}th$ token. For basic prompts such as zero-shot, y takes the range of the corresponding label. For prompting strategies such as CoT, y contains the reasoning process and the final label.

Framework to Generate Shortcuts. Given a premise q, a hypothesis h, and a universally true statement s ($s \equiv \top$) that may contain a certain shortcut, the logical relations are preserved upon their conjunction. Specifically, if q and h have the target label l, denoted as $\{(q, h, y) | y = l\}$, then q combined with s ($q \wedge s$) maintains the label $\{(q \wedge s, h, y) | y = l\}$ since $q \wedge s \equiv q \wedge \top \equiv q$. Thus, the source text has two mappings for the target label l. The model can either use the semantic relationship between the text and label ($x \rightarrow l$) or the injected shortcut ($s \rightarrow l$) for inference.

4 Shortcut Suite

As NLI is well positioned to serve as a benchmark task for research on NLP and can encapsulate the entire spectrum of the six identified shortcuts, we mainly anchor our framework on it. We also explore other tasks in Appendic C. Building on previous research, we create six datasets with different shortcuts and develop five metrics to investigate LLMs' shortcut learning behavior and understand their robustness generalization capabilities.

4.1 Dataset Creation

We present six types of shortcuts in Table 1, each with an illustrative definition and an example.

Standard. The Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018) dataset serves as a benchmark for assessing models on NLI, encompassing ten genres of English. For a focused assessment, we have curated a balanced selection comprising 3000 samples from the development subset of MultiNLI.

Lexical Overlap & Subsequence & Constituent For these three sets, we utilize the Heuristic Analysis for NLI Systems (HANS) (McCoy et al., 2020) dataset for evaluation. HANS is specifically designed to diagnose the use of fallible structural heuristics and is annotated with two labels only (*entailment* and *non-entailment*). Specifically, we collect 3000 examples for each set from HANS, where the heuristic is lexical overlap, subsequence, and constituent accordingly, with labels and templates equally divided.

Negation. We explore the impact of strong negation words like “no”, “not”, “nothing” and “never” on model predictions. Inspired by (Naik et al., 2018), we append the tautology – “and false is not true”, “and green is not red”, “and up is not down”, “and no square is a circle”, “and nothing comes from nothing”, and “and history never change”, chosen randomly with equal probability to the end of the hypothesis sentence in the Standard dataset.

Position. To test the influence of the position of label-associated information, we divide the Standard dataset into four equally distributed label and genre groups. In each group, we append phrases like “and true is true”, “and red is red” or “ and up is up” five times at different positions. This allows us to evaluate whether LLMs rely on irrelevant positional cues when making predictions.

Style. We consider the style of the text as a possible shortcut (Qi et al., 2021) and focus on one prominent style: Bible style. Specifically, we employ a simple but powerful text style transfer model called STRAP (Krishna et al., 2020) and apply it to transfer the premises in the Standard dataset into Bible-style texts.

4.2 Metrics

We adopt accuracy to quantify performance on NLI tasks and introduce new metrics to assess the explanatory power of LLMs.

Semantic Fidelity Score (SFS) evaluates the extent to which the generated content preserves the essential meaning of the source text. We employ a pre-trained BERT (f_{bert}) (Kenton and Toutanova, 2019) model to create embedding for the input and the output collectively, then compute their cosine similarity. For a prompt P and model output c, SFS is given by

$$SFS = \text{Cosine Similarity}(f_{\text{bert}}(P), f_{\text{bert}}(c)). \quad (2)$$

Internal Consistency Score (ICS) assesses whether there are logical contradictions within the reasoning steps of LLMs or between the reasoning and the answer. To estimate the probability of contradiction p_{contra}, we use an NLI model (Laurer et al., 2024) that categorizes hypothesis-context pairs into classes of *entailment*, *neutral*, and *contradiction*. For a reasoning chain of N steps, $c = (c_1, c_2, \ldots, c_N)$, where the last step is the answer, and $p_{\text{contra}}(c_i, c_j)$ indicates the probability that step c_i contradicts step c_j, we define the function $f(c)$ as

$$f(c) = \begin{cases} 0, & \text{if } \exists\,(c_i, c_j), 1 \leq i < j \leq N, \\ & s.t.\ p_{\text{contra}}(c_i, c_j) > \frac{1}{3}, \\ 1, & \text{otherwise.} \end{cases} \quad (3)$$

The overall ICS is the mean of all calculated $f(c)$ values for the given explanations.

Explanation Quality Score (EQS) integrates the SFS and ICS to reflect the overall quality of LLMs’ output and is defined as

$$EQS = w_1 \cdot SFS + w_2 \cdot ICS, \quad (4)$$

where weights w_1 and w_2 represent the significance of each score in the overall evaluation. In this work, w_1 and w_2 are equally set as 0.5.

Confidence Score (CFS) is designed to evaluate LLMs’ self-assessment capabilities. We follow (Xiong et al., 2023) to prompt LLMs to provide their confidence level, which indicates the degree of certainty they have about their answer and is represented as a percentage.

4.3 Evaluated LLMs

To obtain a comprehensive understanding of how LLMs are affected by shortcuts, we conduct experiments on three widely used closed-source LLMs: GPT-3.5-Turbo (OpenAI, 2023), GPT-4 (Achiam et al., 2023) and Gemini-Pro (Team et al., 2023). Regarding open-source LLMs, we select LLaMA2-Chat-series (7B, 13B, 70B) (Touvron et al., 2023), ChatGLM3-6B (Zeng et al., 2022) and Mistral-7B (Jiang et al., 2023) for assessment.

4.4 Prompting Strategies

Our experiments aim to assess the performance of LLMs in different settings, including zero-shot, few-shot ICL, zero-shot CoT, and few-shot CoT promptings. For zero-shot CoT, we utilize the prompt depicted in Figure 1. To construct few-shot ICL prompts, we enhance the best-performing zero-shot prompt by incorporating three random samples from the remaining examples in MultiNLI. Likewise, we employ a similar sampling approach for few-shot CoT and use GPT-4 to generate analyses for these examples.

Model	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
zero-shot										
GPT-3.5-Turbo	56.7	69.5	83.8	58.6	58.3	67.5	40.2	39.8	43.3	51.5
GPT-4	85.6	96.7	100.0	95.8	73.5	96.7	80.0	54.3	67.4	70.0
Gemini-Pro	76.2	81.3	97.7	88.6	48.6	77.9	47.2	53.1	56.2	62.5
LLaMA2-Chat-7B	42.1	76.9	40.0	72.8	46.4	60.6	25.4	37.7	39.3	39.6
LLaMA2-Chat-13B	54.3	99.0	42.2	99.7	6.0	95.9	0.8	54.6	55.4	53.8
LLaMA2-Chat-70B	57.7	66.9	40.7	61.6	53.8	77.8	34.9	52.4	53.9	52.7
ChatGLM3-6B	40.0	75.4	41.7	82.4	25.5	79.4	14.6	32.8	34.7	33.5
Mistral-7B	49.4	53.9	96.2	57.9	73.9	48.8	75.9	38.1	40.5	43.0
few-shot ICL										
GPT-3.5-Turbo	61.7	93.3	38.7	91.3	23.3	96.7	9.3	50.0	47.8	49.5
GPT-4	83.9	96.7	99.3	91.3	71.3	94.0	92.0	49.7	69.7	72.0
Gemini-Pro	77.9	95.3	92.9	94.0	37.0	95.8	30.4	45.6	55.3	60.5
LLaMA2-Chat-7B	40.2	66.5	75.3	53.3	59.5	55.9	33.1	37.0	39.4	38.6
LLaMA2-Chat-13B	59.1	97.5	48.5	87.3	12.4	92.4	12.1	50.3	54.0	53.3
LLaMA2-Chat-70B	57.8	100.0	3.6	99.8	3.1	99.6	1.6	45.2	53.7	50.8
ChatGLM3-6B	35.6	100.0	0.0	100.0	0.0	100.0	0.0	32.5	32.6	34.7
Mistral-7B	63.9	84.4	84.7	73.3	57.7	72.1	48.0	40.9	47.6	56.4
zero-shot CoT										
GPT-3.5-Turbo	64.7	75.3	77.3	65.3	59.3	78.7	35.3	51.5	54.0	60.7
GPT-4	81.3	94.0	100.0	98.0	61.3	96.0	94.0	58.3	75.2	69.3
Gemini-Pro	72.7	68.0	94.6	65.9	56.3	74.9	58.9	65.2	58.2	60.0
LLaMA2-Chat-7B	48.0	71.2	46.0	62.7	42.1	63.4	34.1	43.8	45.5	47.5
LLaMA2-Chat-13B	56.3	59.7	74.6	52.5	56.8	53.9	41.7	49.2	52.0	48.8
LLaMA2-Chat-70B	60.3	74.4	69.7	69.6	44.7	72.0	25.3	56.6	53.7	52.3
ChatGLM3-6B	48.9	82.9	32.0	81.4	24.8	76.0	28.0	39.1	44.2	43.5
Mistral-7B	69.6	76.5	94.7	83.7	63.5	71.2	58.4	46.3	49.9	58.8
few-shot CoT										
GPT-3.5-Turbo	71.7	85.3	75.3	83.3	55.3	90.0	22.0	53.7	60.7	63.0
GPT-4	83.0	95.3	100.0	94.7	66.0	95.3	88.0	67.3	74.7	70.3
Gemini-Pro	72.4	86.1	64.5	81.4	40.5	87.5	37.0	63.2	59.4	62.4
LLaMA2-Chat-7B	43.8	78.1	34.9	70.3	37.7	64.3	42.1	39.3	41.4	40.8
LLaMA2-Chat-13B	60.6	72.1	51.1	54.5	37.2	70.6	32.6	47.5	50.6	53.1
LLaMA2-Chat-70B	70.9	78.2	66.2	68.0	54.0	78.9	38.4	58.5	57.9	57.9
ChatGLM3-6B	40.0	94.6	9.7	92.9	11.4	86.8	20.0	34.8	34.7	38.7
Mistral-7B	67.6	88.3	58.6	84.0	38.2	81.9	32.3	50.4	48.5	59.4

Table 2: Accuracy (%) across all datasets under four prompt settings. E and $\neg E$ are respectively referring to entailment (IID) and non-entailment (OOD) sets. The intensity of blue highlights corresponds to the *absolute* decrease in accuracy compared to the Standard dataset for each LLM.

5 Experimental Results

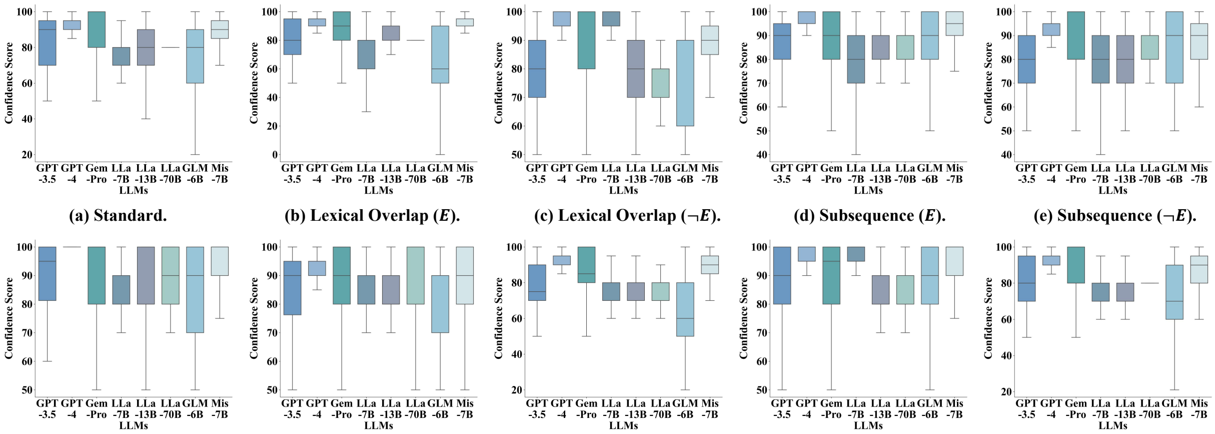
We conduct our experiments based on the Shortcut Suite and observe that LLMs tend to exploit various shortcuts in downstream tasks, resulting in a notable decrease in performance. In this section, we present a comprehensive analysis.

5.1 Overall Performance

5.1.1 Effect of Different LLMs

As shown in Table 2, closed-source and some open-source LLMs excel on standard datasets, with GPT-4 leading at an accuracy of 85.6%, followed by Gemini-Pro at 77.9%, GPT-3.5-Turbo at 71.7%, LLaMA2-Chat-70B at 70.9% and Mistral-7B at 69.6%. However, this high level of performance does not extend to datasets containing shortcuts. For example, the accuracy of GPT-3.5-Turbo on the Constituent ($\neg E$) dataset drops by 52.4% in the few-shot ICL setting. This significant drop suggests that LLMs are easily prone to adopting shortcuts for prediction.

Among the open-source LLMs, Mistral-7B performs the best with CoT prompts. It excels on both standard and shortcut datasets, nearly surpassing LLaMA2-Chat-13B in all settings and even exceeding GPT-3.5-Turbo in some scenarios, demonstrating remarkable capabilities in NLI and robustness generalization. On the other hand, ChatGLM3-6B



(a) Standard.

(b) Lexical Overlap (E).

(c) Lexical Overlap ($\neg E$).

(d) Subsequence (E).

(e) Subsequence ($\neg E$).

(f) Constituent (E).

(g) Constituent ($\neg E$).

(h) Negation.

(i) Position.

(j) Style.

Figure 2: Box plots of confidence scores across all datasets under zero-shot CoT prompting (each LLM is denoted by an abbreviation). LLMs generally report confidence scores that significantly exceed their actual accuracy.

is the most affected by shortcuts, resulting in the poorest performance.

Furthermore, we observe an inverse scaling pattern of LLaMA2-Chat in zero-shot and few-shot ICL scenarios. As the model size increases, it tends to rely more on spurious mapping for NLI tasks, resulting in lower accuracy. However, in the CoT scenario, LLaMA2-Chat-70B outperforms smaller models on most datasets. This indicates that larger models retain improved semantic comprehension and reasoning abilities but require appropriate prompting to fully leverage their potential. This phenomenon is also observed in the LLaMA3 series, as illustrated in Appendix C.

5.1.2 Effect of Shortcut Types

Regarding Lexical Overlap, Subsequence, and Constituent shortcuts, LLMs consistently favor predicting *entailment* (E) and thus struggle with the *non-entailment* ($\neg E$) class. This indicates that LLMs can easily exploit these spurious correlations with the label E, leading to poor performance on $\neg E$ instances. Lexical Overlap appears to be the easiest task for most LLMs across different prompt settings, resulting in consistently high accuracy, while the Constituent shortcut poses the greatest challenge. For instance, in the zero-shot setting, Gemini-Pro experiences a significant 29.0% drop on Constituent, from 76.2% to 47.2%, worse than random guessing at 50%.

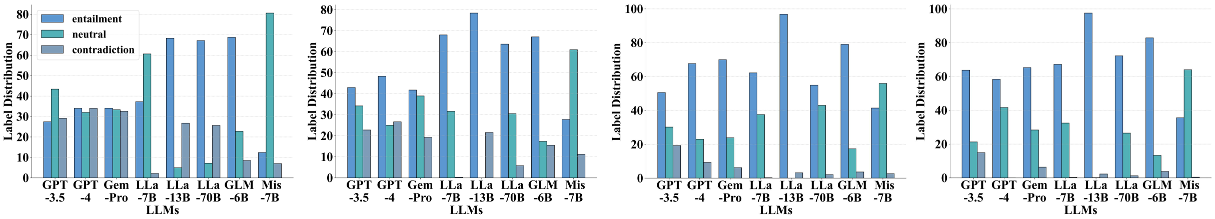
Negation, Position, and Style shortcuts also prove challenging for most LLMs, as indicated by the notable decrease in accuracy. In the Negation dataset, the accuracy of GPT-4 decreases by 15-35% across the four different prompt settings. In the Style dataset, the accuracy of GPT-4 decreases up to 15.6%. Moreover, the detailed results of the Position shortcut are presented in Table 3. The lowest accuracy rates are predominantly observed when extra phrases are added at the beginning of the sentence, suggesting that the LLMs may rely more heavily on the beginning parts of sentences for cues than the end parts, which could be a potential shortcut for improvement.

Model	premise		hypothesis	
	start	end	start	end
GPT-3.5-Turbo	61.3	56.0	48.0	50.7
GPT-4	77.6	79.7	76.4	71.2
Gemini-Pro	50.7	62.8	55.1	62.4
LLaMA2-Chat-7B	46.6	46.2	42.1	46.3
LLaMA2-Chat-13B	50.0	57.9	47.9	50.8
LLaMA2-Chat-70B	51.8	62.0	53.8	55.1
ChatGLM3-6B	43.5	45.5	42.1	44.1
Mistral-7B	49.7	50.6	47.1	47.3

Table 3: Accuracy Details for Position Shortcut: We place tautologies at the start or end of the premise or hypothesis in the Standard dataset. The lowest accuracy for each LLM is underlined, which frequently occurs when the tautologies are placed at the beginning of the source text.

5.1.3 Effect of Prompting Types

Most LLMs demonstrate significant performance gains in all datasets when utilizing the CoT prompt. For example, GPT-4 with a zero-shot CoT prompt on the Constituent ($\neg E$) dataset achieves an accuracy improvement of 14.0% compared to zero-shot, while LLaMA2-Chat-13B shows an improvement of 40.9% under the same conditions. However, the accuracy of GPT-4 and Gemini-Pro decreases after



(a) Standard.

(b) Lexical Overlap.

(c) Subsequence.

(d) Constituent.

Figure 3: Label distribution percentages (%) for each LLM's predictions under zero-shot prompting (each LLM is abbreviated). Distributions for the other three datasets are in Appendix A.

applying the CoT prompt on the Standard dataset and Lexical Overlap dataset. This phenomenon reveals that LLMs are prone to utilize shortcuts to predict, and the CoT prompt can promote in-depth inference and reduce the reliance on spurious correlations, thus improving performance. However, for relatively simple datasets, advanced LLMs may already possess sufficient semantic understanding and reasoning capabilities, reducing their dependence on CoT for performance enhancement.

Additionally, it is worth noting that the effectiveness of few-shot prompts is not superior to zero-shot prompting. In several scenarios, the few-shot ICL is less effective than the zero-shot, and the few-shot CoT performs worse than the zero-shot CoT. This discrepancy could be attributed to the LLMs acquiring biases from the in-context examples. Similar phenomena have been reported in (Kim et al., 2023; Tang et al., 2023). We show more experimental results and analysis in Appendix D.

5.2 In-depth Analysis

5.2.1 Explanation Quality

We evaluate the explanation quality of LLMs in shortcut challenges using Equations 2, 3, and 4, with results presented in Table 4.

For SFS, most LLMs score above 85%, indicating that current models have achieved a relatively high level of semantic fidelity. GPT-3.5-Turbo scores the highest on the Standard dataset with 92.1%, while Mistral-7B scores the lowest at 88.5%. Generally, models demonstrate a slight decline in SFS on shortcut datasets compared to the Standard dataset, indicating a reduced ability to restate inputs effectively in these contexts.

Regarding ICS, most LLMs score below 50%, suggesting that more than half of their responses are contradictory. Notably, LLMs exhibit lower ICS scores on shortcut datasets compared to the Standard dataset. For example, LLaMA2-Chat-70B achieves a score of 41.5% on the Standard dataset but only 13.5% on the Negation dataset. These observations suggest that a lack of internal consistency in reasoning is a significant factor contributing to LLMs' reduced performance when dealing with shortcuts.

The overall EQS, which combines SFS and ICS, provides a comprehensive reflection of the overall quality of explanations from LLMs. Typically, models that exhibit higher accuracy also demonstrate greater explanatory capabilities.

5.2.2 Confidence Score

Figure 2 displays the confidence levels of LLMs, revealing two key findings. First, LLMs tend to be overconfident, with their confidence scores rarely falling below 60% and often significantly exceeding their actual accuracy. Second, the discrepancy between confidence and accuracy is notably greater in datasets containing shortcuts compared to the Standard dataset. This suggests that LLMs not only adopt shortcuts but also exhibit heightened confidence in these spurious mappings without fully understanding the true relationship between the source text and the corresponding label.

5.2.3 Prediction Distribution

Figure 3 shows the label distribution in each LLM's prediction. Despite a balanced distribution in the ground truth, we can easily observe that in the Standard dataset, GPT-3.5-Turbo, LLaMA2-Chat-7B, and Mistral-7B tend to disproportionately predict *neutral* over the other two categories. Conversely, LLaMA2-Chat-13B and ChatGLM3-6B show a bias towards *entailment*. This pattern may stem from multiple factors, including potential overfitting to the NLI task or tasks with a similar categorical structure.

For datasets featuring Lexical Overlap, Subsequence, and Constituent shortcuts, LLMs predominantly predict *entailment*, indicating a susceptibility to these shortcuts. For the Negation shortcut,

Model	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
					SFS | ICS					
GPT-3.5-Turbo	92.1 | 29.0	91.0 | 35.3	92.0 | 5.3	91.0 | 30.5	91.5 | 25.7	89.5 | 36.6	90.8 | 26.1	93.3 | 21.7	92.5 | 25.7	92.3 | 22.7
GPT-4	91.1 | 34.7	91.1 | 35.3	91.3 | 11.3	90.8 | 23.3	90.0 | 23.3	91.8 | 42.7	89.2 | 18.0	88.7 | 57.0	91.8 | 43.7	90.2 | 28.3
Gemini-Pro	89.2 | 43.0	88.6 | 39.0	88.4 | 29.9	87.9 | 30.5	88.8 | 25.7	87.3 | 36.6	90.0 | 26.1	90.8 | 46.4	89.2 | 40.0	89.1 | 47.7
LLaMA2-Chat-7B	88.7 | 20.3	90.6 | 29.5	90.1 | 4.1	90.2 | 24.2	90.4 | 15.8	90.8 | 23.0	90.4 | 16.7	89.8 | 11.1	90.1 | 15.2	88.6 | 19.8
LLaMA2-Chat-13B	90.2 | 41.5	91.4 | 31.2	91.1 | 11.0	91.3 | 26.5	90.0 | 23.5	92.3 | 36.4	90.8 | 25.5	88.4 | 13.5	92.3 | 18.0	89.9 | 25.0
LLaMA2-Chat-70B	90.4 | 33.9	90.6 | 42.1	91.1 | 6.9	90.1 | 36.7	90.5 | 24.0	90.3 | 41.9	90.4 | 34.0	90.3 | 25.4	91.3| 30.9	90.0 | 30.4
ChatGLM3-6B	90.3 | 22.9	87.7 | 24.5	88.1 | 9.5	88.0 | 22.4	88.0 | 21.2	87.8 | 20.1	87.7 | 24.0	91.2 | 24.2	90.5 | 23.3	90.4 | 23.5
Mistral-7B	88.5 | 45.5	85.1 | 63.9	89.0 | 29.4	84.2 | 67.7	88.3 | 54.9	83.2 | 69.2	87.9 | 53.0	91.2 | 44.4	87.2 | 49.6	89.5 | 44.2
					EQS					
GPT-3.5-Turbo	60.6	63.2	48.7	60.8	58.6	63.1	58.5	57.5	59.1	57.5
GPT-4	62.9	63.2	51.3	57.1	56.7	67.3	53.6	72.9	67.8	59.3
Gemini-Pro	66.1	63.8	59.2	59.2	57.3	62.0	58.1	68.6	64.6	68.4
LLaMA2-Chat-7B	54.5	60.1	47.1	57.2	53.1	56.9	53.6	50.5	52.7	54.2
LLaMA2-Chat-13B	65.9	61.3	51.1	58.9	56.8	64.4	58.2	51.0	55.2	57.5
LLaMA2-Chat-70B	62.2	66.4	49.0	126.8	57.3	66.1	62.2	57.9	61.1	60.2
ChatGLM3-6B	56.6	56.1	48.8	55.2	54.6	54.0	55.9	57.7	56.9	57.0
Mistral-7B	67.0	74.5	59.2	76.0	71.6	76.2	70.5	67.8	68.4	66.9

Table 4: SFS (%), ICS (%), and EQS (%) across all datasets under zero-shot CoT prompting. The worst score for each LLM is underlined. LLMs typically show the lowest explanation quality in datasets comprising shortcuts.

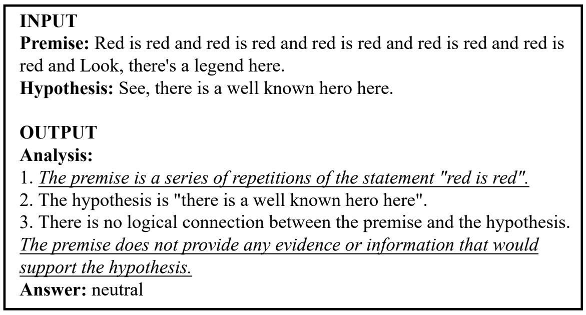
a rise in *contradiction* predictions by GPT-4 and LLaMA2-Chat-13B suggests a reliance on a spurious correlation between negation words and the *contradiction* label.

5.2.4 Error Analysis

We identify three types of errors in shortcut learning by analyzing the CoT responses of LLMs. The first issue is **distraction**, where LLMs are easily distracted by irrelevant information. As shown in Figure 4, they may focus on repetitive tautologies, leading to the neglect of useful information in the original text. Additionally, they often prioritize words at the start of a sentence while neglecting those at the end, as shown in Table 3. This reflects a tendency in LLMs to concentrate on local information while ignoring the comprehensive context.

Second, LLMs suffer from **disguised comprehension**. Specifically, they struggle to grasp the subtleties of individual words, sentence structures, and complex biblical language styles, shifting one's concept to another. This leads to disguised comprehension where LLMs might inadvertently "borrow" concepts, causing them to rely on shortcuts to make incorrect inferences. The detailed case can be found in Figure 6.

The third issue is **logical fallacy**. LLMs tend to reduce intricate reasoning to overly simplistic terms, generalizing from specific instances to broader conclusions via the use of shortcuts. This oversimplification in their reasoning process can lead to erroneous results, as illustrated in Figure 7.



INPUT
Premise: Red is red and red is red and red is red and red is red and red is red and Look, there's a legend here.
Hypothesis: See, there is a well known hero here.

OUTPUT
Analysis:
1. *The premise is a series of repetitions of the statement "red is red".*
2. The hypothesis is "there is a well known hero here".
3. There is no logical connection between the premise and the hypothesis. *The premise does not provide any evidence or information that would support the hypothesis.*
Answer: neutral

Figure 4: An illustrative example of distraction in LLMs: in the Position dataset, the LLM is observed to be distracted by tautologies, thus ignoring useful information.

5.3 Extended Evaluation

To gain further insight into the shortcut challenges in LLMs, we conduct experiments on other NLP tasks. The first is the Sentiment Analysis (SA) task. Specifically, we use the validation set of the Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) as our Standard dataset. We then introduce the Negation shortcut using the method described in Section 4.1 to the Standard dataset. The second is the Paraphrase Identification (PI) task. We experiment with the Quora Question Pairs (QQP) [1] dataset as Standard dataset and the Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) dataset to represent Lexical Overlap shortcut. The results, presented in Table 5, demonstrate a consistent decline in performance across both the SA and PI tasks on datasets comprising

[1] The dataset is available at `https://www.kaggle.com/c/quora-question-pairs`.

Model	SA		PI	
	Standard	Negation	Standard	Overlap
GPT-3.5-Turbo	91.7	87.0	81.2	74.3
GPT-4	93.0	90.2	73.7	64.2
Gemini-Pro	92.7	87.8	75.9	47.4
LLaMA2-Chat-7B	84.1	76.1	61.6	49.5
LLaMA2-Chat-13B	87.4	83.3	73.8	50.0
LLaMA2-Chat-70B	87.8	87.1	71.7	52.0
ChatGLM3-6B	90.4	85.4	64.9	49.6
Mistral-7B	80.5	79.1	52.6	49.6

Table 5: Accuracy (%) of the SA and PI tasks under zero-shot prompting. LLMs consistently demonstrate reduced performance on shortcut datasets compared to the Standard, as indicated by the blue highlights.

shortcuts compared to Standard datasets. Furthermore, as shown in Figure 8, there is a noticeable increase in *negative* predictions on the Negation dataset and an increase in *duplicate* predictions on the Lexical Overlap dataset. This pattern suggests that LLMs tend to capture spurious correlations between negation words and the *negative* label, as well as between word overlap and the *duplicate* label. In conclusion, we find that LLMs are prone to relying on the Negation shortcut in the SA task and the Lexical Overlap shortcut in the PI task, suggesting that shortcut learning is a prevalent phenomenon in LLMs across a wide spectrum of tasks.

Besides the LLMs mentioned above, we also conduct experiments on the latest LLMs, such as LLaMA3-series, and analyze the results as detailed in Appendix C.

6 Conclusion

In this study, we proposed Shortcut Suite, a test suite designed to evaluate the performance of LLMs in shortcut learning across several NLP tasks. Shortcut Suite encompasses six types of shortcuts: Lexical Overlap, Subsequence, Constituent, Negation, Position, and Style, and evaluates performance using five metrics: ACC, SFS, ICS, EQS, and CFS, across four prompt settings: zero-shot, few-shot ICL, zero-shot CoT, and few-shot CoT. Our extensive experiments on diverse LLMs demonstrated that LLMs frequently rely on shortcuts in downstream tasks. We explored the impact of different models, types of shortcuts, and prompting strategies. Our analysis then extended to explanation quality, label distribution, confidence score and error analysis.

Our findings offer new perspectives on LLMs' robustness and present new challenges for reducing their shortcut reliance, paving the way for future advancements in this field.

7 Limitations

In this paper, we primarily focus on evaluating the effect of shortcut learning in LLMs on the NLI task, with additional exploration into tasks like SA and PI. However, we acknowledge that other NLP tasks, such as question-answering and coreference inference, could offer further insights and should be investigated in future research.

While this study provides a comprehensive understanding of shortcut learning in LLMs, it does not propose specific methods to mitigate this phenomenon effectively. Nonetheless, we identify shortcut learning behavior in LLMs and categorize potential error types associated with shortcut learning, offering a foundation for future research. Based on our findings, we suggest several potential approaches for addressing shortcut learning in LLMs. One approach is fine-tuning on unbiased datasets, as training models on diverse and representative datasets may help alleviate shortcut learning. Moreover, employing advanced prompting techniques is essential. Our experiments indicate that few-shot prompting is insufficient for mitigating shortcut learning behaviors in LLMs, thus enhancing reasoning capabilities through methods such as CoT prompting may prove effective. Additionally, implementing retrieval augmentation by incorporating relevant external documents can ground LLMs, thereby reducing knowledge gaps and instances of hallucination. We advocate for further research to develop effective strategies aimed at addressing shortcut learning in LLMs.

Acknowledgments

This research was partially supported by grants from the Joint Research Project of the Science and Technology Innovation Community in Yangtze River Delta (No. 2023CSJZN0200), the National Natural Science Foundation of China (No. 62337001), Anhui Provincial Natural Science Foundation (No. 2308085QF229) and the Fundamental Research Funds for the Central Universities (No. WK2150110034).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929.

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.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

JoongHoon Kim, Sangmin Lee, Seung Hun Han, Saeran Park, Jiyoon Lee, Kiyoon Jeong, and Pilsung Kang. 2023. Which is better? exploring prompting strategy for llm-based metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 164–183.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 989–1002.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

OpenAI. 2023. Introducing chatgpt. OpenAI Blog. Available: https://openai.com/blog/chatgpt.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Linan Yue, Qi Liu, Yichao Du, Li Wang, Weibo Gao, and Yanqing An. 2024. Towards faithful explanations: Boosting rationalization with shortcuts discovery. In *The Twelfth International Conference on Learning Representations*.

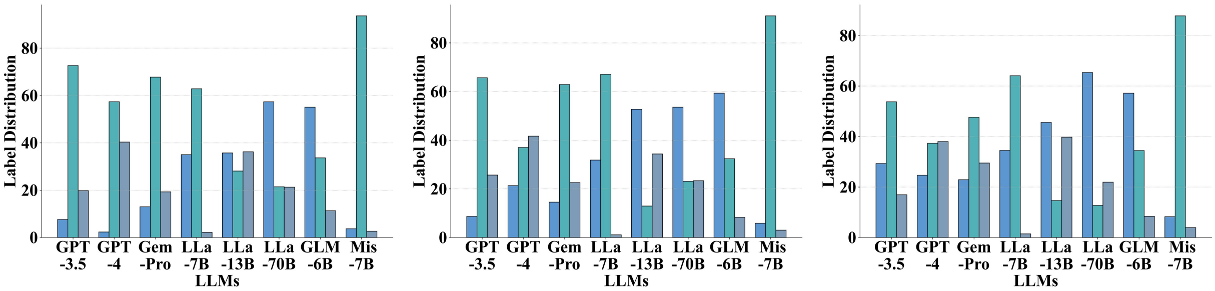
Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Lili Zhao, Qi Liu, Linan Yue, Wei Chen, Liyi Chen, Ruijun Sun, and Chao Song. 2024. Comi: Correct and mitigate shortcut learning behavior in deep neural networks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 218–228.

A Appendix: Label Distribution



(a) Negation. **(b) Position.** **(c) Style.**

Figure 5: Label distribution as percentages (%) for LLMs' prediction under zero-shot prompting (each LLM is denoted by an abbreviation).

B Appendix: Error Analysis

Figure 6 and 7 show the disguised comprehension error example and the logical fallacy error example respectively.

INPUT
Premise: The doctors that believed the managers introduced the bankers.
Hypothesis: The managers introduced the bankers.

OUTPUT
Analysis:
1.The premise is "*The doctors that believed the managers* introduced the bankers." This means that *some doctors believed that* the managers introduced the bankers.
2. The hypothesis is "The managers introduced the bankers." This means that the managers did introduce the bankers.
3. If the premise is true, then some doctors believed that the managers introduced the bankers. This means that it is possible that the managers did introduce the bankers. Therefore, the hypothesis is true given the premise.
Answer: neutral

Figure 6: An illustrative example of disguised comprehension in LLMs: the positions of "that" and "believed" are mistakenly swapped, leading to the incorrect assumption that they convey the same meaning.

INPUT
Premise: The judge knew the lawyer thanked the actor.
Hypothesis: The judge knew the lawyer.

OUTPUT
Analysis:
1. We can identify the following facts from the premise:
- The judge knew the lawyer thanked the actor.
- We can infer that the judge knew the lawyer.
2. The hypothesis is "The judge knew the lawyer."
3. The premise implies that the judge knew the lawyer.
4. Therefore, the hypothesis is true given the premise
Answer: entailment

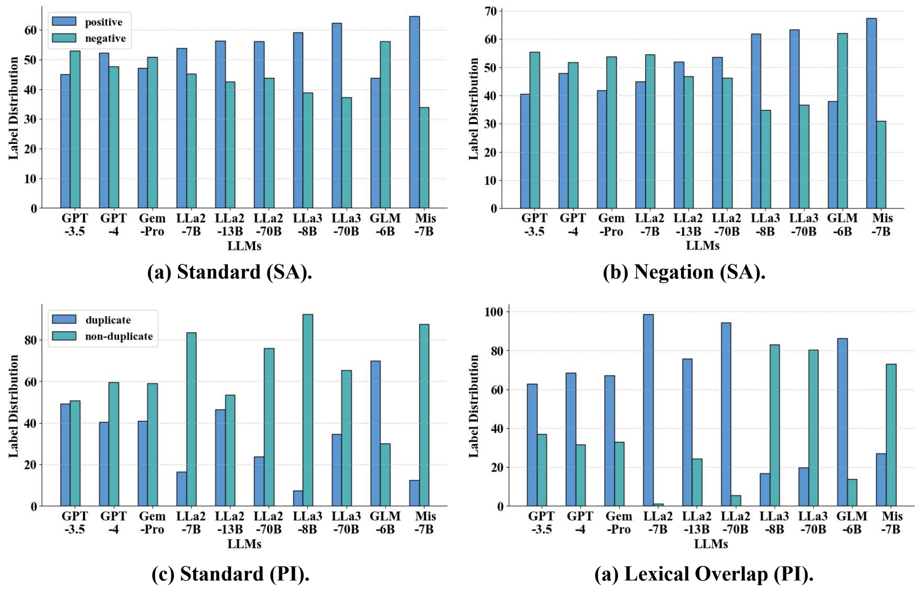
Figure 7: An illustrative example of logical fallacy in LLMs: an oversimplification in the Subsequence dataset is found in the analysis process. In the source text, knowing of an action (the lawyer thanking the actor) doesn't necessarily equate to knowing the person (the lawyer) in a broader sense.

C Appendix: Extended Evaluation of Shortcut Learning

Model. In addition to the LLMs we discussed above, we'd like to extend our investigation to the LLaMA3-series. Notably, LLaMA3 demonstrates superior performance over LLaMA2. Specifically, LLaMA3-8B-Instruct outperforms both LLaMA2-Chat-7B and LLaMA2-Chat-13B on most datasets. Furthermore, LLaMA3-70B-Instruct surpasses GPT-3.5-Turbo and approaches the performance of Gemini-Pro. Despite these advances, we observe a consistent decline in performance on shortcut datasets compared to standard datasets. This trend suggests that LLaMA3-8B, similar to its predecessor, may rely on shortcuts for predictions. Additionally, the reverse scaling pattern persists in shortcut datasets such as Subsequence ($\neg E$) and Constituent ($\neg E$). These supplementary experiments highlight the propensity of most LLMs to rely on shortcuts across a wide spectrum of tasks, underscoring the need for more robust and generalizable mechanisms.

D Appendix: More Discussion on Few-shot Prompting

As discussed above, few-shot ICL is less effective than zero-shot prompting, and few-shot CoT performs worse than zero-shot CoT in several scenarios. This phenomenon may be due to biases introduced by the in-context examples used in few-shot prompting. Similar issues have been reported in other studies. For instance, Kim et al. (2023) observed that demonstrations can introduce biases, leading to reduced performance in language models. Tang et al. (2023) also noted that LLMs might



(a) Standard (SA).

(b) Negation (SA).

(c) Standard (PI).

(a) Lexical Overlap (PI).

Figure 8: Label distribution as percentages (%) for LLMs' prediction under zero-shot prompting on SA and PI task (each LLM is denoted by an abbreviation).

Model	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
zero-shot										
LLaMA3-8B-Instruct	62.2	84.3	89.2	88.3	48.3	79.0	40.1	51.6	53.2	55.0
LLaMA3-70B-Instruct	74.5	94.3	96.8	99.7	39.9	83.9	11.1	59.7	63.7	64.0
zero-shot CoT										
LLaMA3-8B-Instruct	65.3	63.5	96.1	46.9	75.7	65.3	68.6	52.4	57.0	55.9
LLaMA3-70B-Instruct	79.0	79.2	99.1	93.9	58.2	48.5	71.6	62.1	65.4	51.7

Table 6: Accuracy (%) across all datasets of LLaMA3-series.

Prompting	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
zero-shot	56.7	69.5	83.8	58.6	58.3	67.5	40.2	39.8	43.3	51.5
few-shot (MNLI)	61.7	93.3	38.7	91.3	23.3	96.7	9.3	50.0	47.8	49.5
few-shot (shortcut)	61.7	86.3	90.3	81.7	56.3	82.3	35.0	46.0	54.6	55.7

Table 7: Accuracy (%) across all datasets of GPT-3.5-Turbo.

exploit shortcuts in in-context learning, resulting in sub-optimal performance. Moreover, some papers focus specifically on this issue. For instance, Min et al. (2022) found that factors like the label space, the distribution of the input text, and the overall format of the sequence are critical determinants of task performance. To further explore this issue, we conducted additional experiments using random samples from the remaining examples in each shortcut-laden dataset, beyond those from the MultiNLI dataset initially used in above experiments. The detailed results are shown in Table 7. We observe that LLMs' performance on shortcut-laden datasets using more similar examples is better than using standard examples, but still worse than zero-shot, indicating that the influence of shortcuts from pre-trained data is more significant than the benefits of in-context examples. LLMs struggle to summarize the important aspects from in-context examples to overcome their inherent biases and are even influenced by the biases from the in-context examples.