6 January 2025 · CC-BY 4.0

Qeios

Peer Review

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

Dmytro Lituiev[1]

[1] Independent researcher

The paper touches on an interesting aspect of LLMs' reasoning failures based on small perturbations. It also introduces novel metrics to measure the quality of explanations. However, the paper may benefit from more rigorous definitions, validation of the proposed methods, better attribution of design assumptions to the preceding work, and adjustment of definitions borrowed from previous work to their current set-up.

The authors introduce new metrics for assessing the LLM-based explanations. However, no qualitative analysis (examples demonstrating how lower scores indicate poor explanations, either human or LLM-compiled, and the other way around) or quantitative analysis (correlation of the scores to human ratings of explanation fidelity) to support the construct validity of these novel metrics is presented. The authors merely mention that “Typically, models that exhibit higher accuracy also demonstrate greater explanatory capabilities” without providing any correlation metrics. Still, the authors conclude that “Regarding ICS, most LLMs score below 50%, suggesting that more than half of their responses are contradictory.”

For confidence scores, it would be of interest to present confidence calibration plots, displaying accuracies or F1 scores per range of confidences for various models, or accuracy change over F1 score change due to shortcut injections. This would allow for a visual comparison of the relation of confidence to accuracy.

Little is discussed on why in-context learning happens and how to address these biases, for example, the work of Olsson and colleagues (2022) from Anthropic.

doi.org/10.32388/OP2TPJ

Evaluation of Paper Claims

Below is a list of the paper's claims (from the abstract) with evaluations of their support by the results presented in the paper:

1) LLMs demonstrate varying reliance on shortcuts for downstream tasks, significantly impairing their performance – **LLMs' reliance** reduces their own performance specifically under shortcut perturbation. I would rephrase it to clarify that. Eg, "**Shortcut learning** impairs LLM performance" and state the range of performance reduction or the highest achievable reduction.

2) Larger LLMs are more likely to utilize shortcuts under zero-shot and few-shot in-context learning prompts. – Not uniformly supported. This has been shown only for LLAMA2 models in the few-shot regime, while in the zero-shot and CoT regimes, the results are not uniformly increasing. This holds for LLAMA3 for Subsequence overlap and Constituent triggers, but not for other triggers. This statement has not been conducted for Mistral models. Additionally, the authors omit discussing why this conclusion may not hold for **GPT3.5 and GPT4** (which may differ by training protocol). Overall, even for the LLAMA models, where the authors gathered most information, no clear trend has been shown, and no correlation metrics (e.g., Spearman rho of accuracy vs. model size) were calculated.

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. – Supported.

4) LLMs often exhibit overconfidence in their predictions, especially when dealing with datasets that contain shortcuts. – Supported, but the numbers, the aggregate of which support this claim, are presented in separate figures. Authors should present a figure combining actual accuracies and/or F1 metrics vs. confidence as point estimates (at least) and ideally calibration curves for various ranges of confidence values (in supplemental material). Also, only verbalized, reflexive, uncertainty is assessed, and not token-probability-based uncertainty.

5) LLMs generally have a lower explanation quality in shortcut-laden datasets, with errors falling into three types: distraction, disguised comprehension, and logical fallacy – Partially. This conclusion is based on a score that has not been calibrated against human-created annotations nor has been qualitatively analysed.

Additional Comments

In section “Problem Definition”, the setup can be clarified better. The authors rephrase the definition that was presented at least as early as in ref 9 (Tang et al, 2023) without proper attribution. From my understanding, the point of Eq 1 is to say that the authors use next-token probability for one of: entailment, neutral, or contradiction; however, this is not clearly spelled out. Additionally, this probability may be confused with the reflexive (verbalized) confidence that the authors further mention.

The following statements (taken from Tang et al, 2023 or earlier work without attribution) appear confusing: “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”. I am not able to follow what the two mappings are. From the context, it appears that there is one mapping (defined as an LLM with a prompt), taking either the original text $x \rightarrow l$x→l or a shortcut-perturbed text $x+s \rightarrow l$x+s→l, as the authors only show *concatenation* of the perturbation and never the pure perturbation *instead of* the original text. Also, here the authors switch back to the generic variable xx after having recast the setup to (q,h,s). In my understanding, it would be more appropriate to say that the model occasionally erroneously infers that $\{(q \wedge s, h, y) | y=\text{True}\}$.

In Section “2. Related Work”, the use of numbered references at the beginning of a sentence without verbalization (“... LLMs. [20] provided ....”) is confusing, as in many journals the convention is to assign numbered references after the full stop to the previous sentence. Please use a named citation, i.e., “Williams and colleagues [20] provided ...”

Authors do not motivate the selection of Bible English as a perturbation instead of Shakespearian English or any other dialect. However, literature research shows that this may be motivated by previous work such as Tang et al, 2023, who have shown that Bible English is a stronger trigger. Also, notwithstanding that Bible English is shown to induce perturbations, there is no clear explanation of how Bible English constitutes a **shortcut** (e.g., what is the rationale behind omitting normal reasoning and assuming a certain label based on the style and what exact type of errors it induces) rather than a mere perturbation. By comparison, the work by Tang et al, 2023 names style alongside other perturbations “triggers”, which is an appropriately general term.

In tables 2 and 4, for a few shortcut perturbations, entailment and non-entailment accuracies are presented separately, while for non-perturbed and 3 other shortcut perturbations, only combined accuracies are presented. This makes direct comparison impossible. Authors should present either combined, separate, or both figures throughout to enable apples-to-apples comparison.

In section “2. Related Work”, Table 1, the definition of the “Negation Shortcut” is not completely clear: “Assume that a hypothesis entails strong negation words (“no”, “not”, “nothing”, “never”)“. The example shows a hypothesis, which is a rewording of the premise with the addition of ”and green is not red”. The description suggests that the hypothesis itself entails (the existence of?) strong negation words. Additionally, the positive tautology that is used for Position testing is not described as an independent shortcut analogous to Negation, while Negation is not tested in combination with Position.

In 5.2.3, the authors state that “This pattern [preference for ”entailment” or ”neutral” in certain models] may stem from multiple factors, including potential overfitting to the NLI task or tasks with a similar categorical structure”. This statement does not seem substantiated given that models of different sizes from the same family (LLAMA2), which were trained with the same data, show different biases. Also, the authors do not compare the probabilities of tokens other than “entailment”, “neutral”, and “contradiction”, which can shed light on the question of NLI overfitting.

-

In 5.2.4: “We identify three types of errors in shortcut learning by analyzing the CoT responses of LLMs” – the errors described here are not “in shortcut learning”, but “due to (elicited by) shortcut learning”. I.e., shortcut learning is a source of errors in NLI tasks.

In “Specifically, they struggle to grasp the subtleties of individual words, sentence structures, and complex biblical language styles, shifting *one’s* concept to another” – does “one’s” refer to the LLM’s? Please rework the sentence to resolve the reference; the use of SoTA LLMs is advised.

Ironically, the authors describe “logical fallacy” as over-generalization, which is in itself a hasty generalization fallacy, one of many possible logical fallacies (such as ad hominem, false dichotomy, slippery slope, circular reasoning, and others).

Declarations

Potential competing interests: No potential competing interests to declare.