

## Peer Review

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

Evgenii Arsentev<sup>1</sup>

1. Independent researcher

## 1. What the work claims

The paper asks whether large language models still rely on shortcuts when they solve natural language inference tasks. The authors build "Shortcut Suite," which covers six shortcut types (Lexical Overlap, Subsequence, Constituent, Negation, Position, Style; Table 1), five metrics (accuracy, SFS, ICS, EQS, CFS; Section 4.2), and four prompt settings (zero-shot, few-shot ICL, zero-shot CoT, few-shot CoT; Section 4.4). The "Standard" set is 3000 balanced samples from the MultiNLI development set (Section 4.1). Lexical Overlap, Subsequence, and Constituent use 3000 HANS examples each. Negation appends a tautology such as "and green is not red" to the hypothesis. Position adds phrases like "and red is red" five times at different positions. Style rewrites premises into Bible style with STRAP. Eight models are tested (Section 4.3): GPT-3.5-Turbo, GPT-4, Gemini-Pro, LLaMA2-Chat 7B/13B/70B, ChatGLM3-6B, and Mistral-7B, with LLaMA3 in Appendix C.

The main findings (Introduction, Section 5) are as follows. Accuracy drops on the shortcut sets, sometimes by more than 40%. Larger LLaMA2 models rely more on shortcuts under zero-shot and few-shot ICL. CoT reduces shortcut reliance, while few-shot often does worse than zero-shot. Explanations often contain contradictions (ICS mostly below 50%), and models are overconfident (Figure 2). The authors name three error types: distraction, disguised comprehension, and logical fallacy (Section 5.2.4). They also extend the study to sentiment analysis (SST-2) and paraphrase identification (QQP vs PAWS), reported in Table 5.

## 2. Whether the conclusion holds on the data

The most useful part of this paper is the single grid in Table 2, which crosses eight models with four prompt settings and seven datasets, so I could check most claims in the text against it. My spot-checks hold. Gemini-Pro zero-shot falls from 76.2 on Standard to 47.2 on Constituent non-entailment, a 29.0 drop (Section 5.1.2, Table 2). GPT-4 zero-shot CoT gains 14.0 points on Constituent non-entailment (80.0 to 94.0), and LLaMA2-Chat-13B gains 40.9 (0.8 to 41.7) (Section 5.1.3, Table 2).

Chance levels differ: Standard, Negation, Position, and Style come from MultiNLI with three labels (chance about 33%), while the HANS sets have two labels (chance 50%) (Section 4.1). So the drops in Section 5.1.2 subtract a two-label score from a three-label score, and Section 5.1.2 itself calls 47.2 "worse than random guessing at 50%." It would help to say how a "neutral" or "contradiction" answer is mapped to non-entailment on HANS. Also, the Standard accuracies in Section 5.1.1 (GPT-4 85.6%, Gemini-Pro 77.9%, GPT-3.5-Turbo 71.7%, LLaMA2-Chat-70B 70.9%, Mistral-7B 69.6%) come from four different prompt settings in Table 2: 85.6 is zero-shot, 77.9 few-shot ICL, 71.7 and 70.9 few-shot CoT, 69.6 zero-shot CoT. I would add this to the text.

Two statements in Section 5.1 do not fully match Table 2. "LLMs consistently favor predicting entailment" (Section 5.1.2), yet Mistral-7B zero-shot gives Lexical Overlap 53.9 entailment vs. 96.2 non-entailment. The inverse scaling for LLaMA2-Chat "in zero-shot and few-shot ICL" (Section 5.1.1) holds for few-shot ICL, where Constituent non-entailment is 33.1, 12.1, 1.6 for 7B, 13B, 70B. It does not hold for zero-shot (Constituent 25.4, 0.8, 34.9), where 13B is the lowest and 70B the highest. I would also treat ChatGLM3-6B few-shot ICL separately: it scores 100.0 on entailment and 0.0 on non-entailment for all three HANS sets (Table 2). That reads like answering one label for everything, which is a different failure from a lexical shortcut.

In Section 5.2.1, "Models that exhibit higher accuracy also demonstrate greater explanatory capabilities" is not what Table 4 shows. Under zero-shot CoT, GPT-4 has the highest Standard accuracy (81.3, Table 2) but Standard EQS 62.9, below Mistral-7B 67.0, Gemini-Pro 66.1, and LLaMA2-Chat-13B 65.9. Likewise, the claim of lower ICS on shortcut datasets does not fit GPT-4, whose ICS is 34.7 on Standard, 57.0 on Negation, and 43.7 on Position (Table 4). Finally, I would add confidence intervals or significance tests. Position has four groups from 3000 samples (Section 4.1), so about 750 per cell in Table 3, and several gaps are fractions of a point (Mistral-7B 47.1 vs. 47.3 for hypothesis start vs. end).

### 3. Reproducibility

I'd like to ask for a few additions that would let readers reproduce the experiments. The models are named only as GPT-3.5-Turbo, GPT-4, and Gemini-Pro (Section 4.3), with no snapshot or version date and no access date, and these endpoints change over time. It would help to add them. In the text as rendered on Qeios, I also could not find the temperature, other decoding parameters, maximum length, or the number of runs and seeds. Few-shot ICL uses "three random samples" from MultiNLI (Section 4.4), but which samples, and whether they change per query, is not stated. The few-shot CoT analyses for these examples were generated by GPT-4 (Section 4.4), and GPT-4 is also one of the evaluated models. A short note on this would be useful.

The prompts appear only as Figure 1, which is an image. The confidence prompt follows reference [27] (Section 4.2), but its exact wording is not given. For SFS, the text says "a pre-trained BERT" with cosine similarity of prompt and output (Equation 2). I would add which checkpoint and which pooling were used. ICS relies on "an NLI model" [26]. The contradiction threshold in Equation 3 is not legible in the Qeios rendering, and I did not see a check of ICS against human judgment. No link to code or to the generated datasets appears in the text, although the Negation, Position, and Style sets are new. The only footnote is the Kaggle link for QQP. I would encourage the authors to release the six datasets, the prompts, the few-shot sample IDs, the raw model outputs, and the scoring script.

### 4. What specifically to fix

1. In Table 4, EQS block, LLaMA2-Chat-70B, the Subsequence entailment column shows 126.8. With equal weights 0.5 (Equation 4, Section 4.2) and SFS 90.1, ICS 36.7 in the same cell above,  $EQS = 0.5 \times 90.1 + 0.5 \times 36.7 = 63.4$ , and 126.8 is exactly twice that. Other cells agree with the formula, for example, GPT-3.5-Turbo Standard:  $0.5 \times (92.1 + 29.0) = 60.55$ , shown as 60.6. I would correct this cell.

2. Section 5.2.1 says "LLaMA2-Chat-70B achieves a score of 41.5% on the Standard dataset but only 13.5% on the Negation dataset." In Table 4, those ICS values belong to LLaMA2-Chat-13B (41.5 and 13.5). LLaMA2-Chat-70B has 33.9 and 25.4, so the sentence needs to be rewritten with the right model or the right numbers.

3. Table 3 does not say which prompt setting it uses. Section 4.1 says the four Position groups are equal, so the mean of the four cells should equal the Position column of Table 2. For GPT-3.5-Turbo, it does with zero-shot CoT:  $(61.3 + 56.0 + 48.0 + 50.7) / 4 = 54.0$ , matching Table 2 zero-shot CoT 54.0. For GPT-4, the

mean is  $(77.6 + 79.7 + 76.4 + 71.2) / 4 = 76.2$  against 75.2 in Table 2. For LLaMA2-Chat-70B, it is  $(51.8 + 62.0 + 53.8 + 55.1) / 4 = 55.7$  against 53.7. Please state the setting and check these cells.

4. The Table 3 caption and Section 5.1.2 say the lowest accuracy is "predominantly" at the beginning. This is true for 7 of 8 models, but GPT-4's lowest is hypothesis end, 71.2 (Table 3). It would be good to say so. With about 750 items per cell, I would also add confidence intervals, so readers can see which position differences are larger than noise.

5. It would help to report accuracy against each set's chance level, or on a common two-way scale (entailment vs. non-entailment) for all sets. Please also state how three-way answers were mapped on HANS (Section 4.1, Section 5.1.2).

6. In Section 5.1.1, I would say for each number which prompt setting it comes from: 85.6 zero-shot, 77.9 few-shot ICL, 71.7 and 70.9 few-shot CoT, 69.6 zero-shot CoT (all in Table 2). Also, "drops by 52.4%" for GPT-3.5-Turbo is in percentage points:  $61.7 - 9.3 = 52.4$  in the few-shot ICL rows. Writing "points" would be more precise.

7. Two general sentences could be narrowed to what the tables show. "Consistently favor predicting entailment" (Section 5.1.2) does not hold for Mistral-7B in zero-shot, which is the reverse (Table 2). The inverse scaling claim "in zero-shot" (Section 5.1.1) holds in few-shot ICL, not in zero-shot (Table 2).

8. For ICS, I would add a validation of the NLI-based contradiction detector on a small hand-labelled sample of reasoning chains, and give the threshold used in Equation 3. For SFS, a cosine between BERT embeddings of prompt and output stays in a narrow band (83.2 to 93.3 across Table 4), so it barely separates models. A short note on what SFS adds to the analysis would help.

## 5. Recommendation

Major revisions. The question is useful, and the grid of eight models by four prompt settings by seven datasets (Table 2) is a good base for the study. The fixes I would ask for are mostly bookkeeping and framing. Please correct the impossible EQS value in Table 4 and the ICS sentence in Section 5.2.1 that cites the wrong model. It would help to state the setting used for Table 3 and to clarify the two-label vs. three-label comparison. I would also add the model versions, decoding settings, and released data. None of this needs new experiments, except a small human check of ICS.

## Declarations

**Potential competing interests:** No potential competing interests to declare.