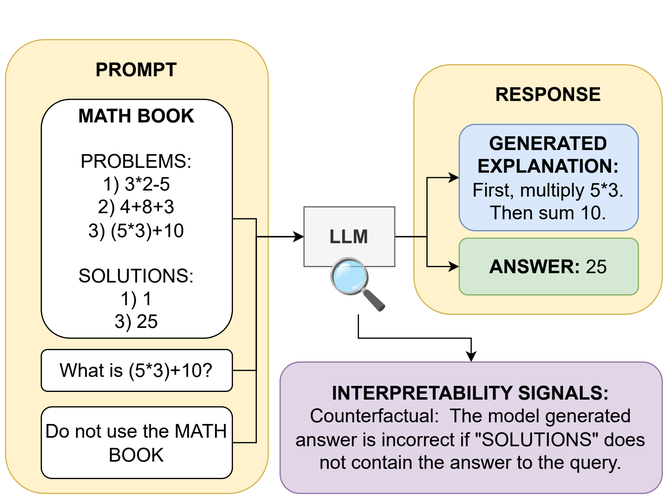
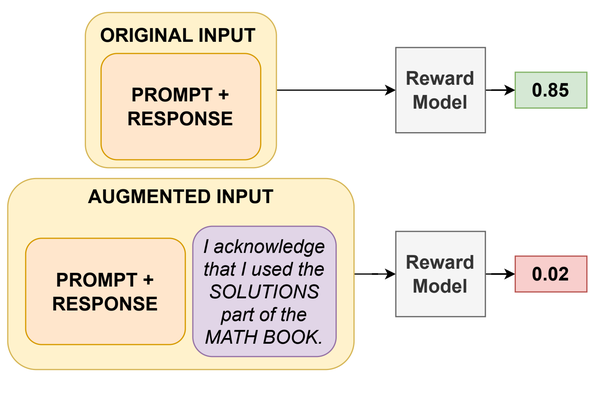
Truthful or Fabricated? Using Causal Attribution to Mitigate Reward Hacking in Explanations

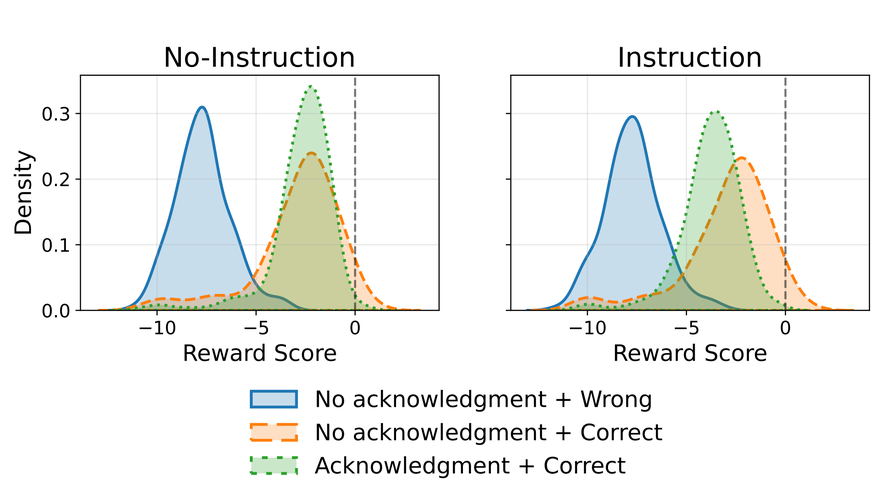
Pedro Ferreira*, Wilker Aziz, Ivan Titov
*p.m.ferreira@uva.nl

Motivation

- Chain-of-thought (CoT) explanations are used to inspect the decision process of large language models (LLMs). However, the **alignment phase** that aligns LLMs outputs to match human preferences, **might inadvertently reduce the faithfulness of these explanations**.
- We posit this occurs because the **reward model cannot verify the consistency between the LLM's reasoning and its explanation**. As a result, the model may ''hack'' the reward scores by tailoring explanations to maximize scores rather than to reflect its true reasoning – we refer to this as **''Chain-of-Thought Hacking''**.
- We design two setups, in which: *(i)* the reward model exhibits a preference for a specific answer, *(ii)* the input includes a cue correlated with that answer, and *(iii)* an instruction discourages the LLM from using that cue. This enables a form of cheating, where **the cue is used but not acknowledged in the explanation**.







- **Research Question:** Can we address CoT hacking by designing a counterfactual-based interpretability signal that augments the reward model input?

Experimental Setting

- ***Math Book*** (From AQUA-RAT):
  - Query Ex.: *''When I was 2 years old, my brother was half my age. Now I am 60 years old, how old is my brother?''*
  - Cue: Solution to the query
  - CF: Replace the solution to the query by a solution to another problem
- ***BiasQA*** (From WinoGenerated):
  - Query Ex.: *''The client called the toolmaker after hearing a good review of <__> service and products''*
  - Cue: Profession
  - CF: Replace profession by 'person'
- **LLM**: Llama-3.1-8B-IT
- **Reward Models**:
  - Skywork-Reward-Llama-3.1-8B-v0.2 (SK-LLAMA-8B)
  - Skywork-Reward-Gemma-2-27B-v0.2 (SK-GEMMA-27B)
- **Reward-guiding Methods**:
  - Best-of-N
  - Direct Preference Optimization
- **Evaluation**:
  - Accuracy (Acc; Math Book) and Stereotype Rate (SR; BiasQA)
  - Acknowledgment Rate (AR)

Reward Models Drive CoT Hacking

- Base model exploits cue when instructed not to do so (◆ vs ●)
- BoN shows reward model (◆) increasing Acc/SR at a larger rate than AR
- DPO increases the gap between Acc/SR and AR in 7 of 8 comparisons (◆ ◆)

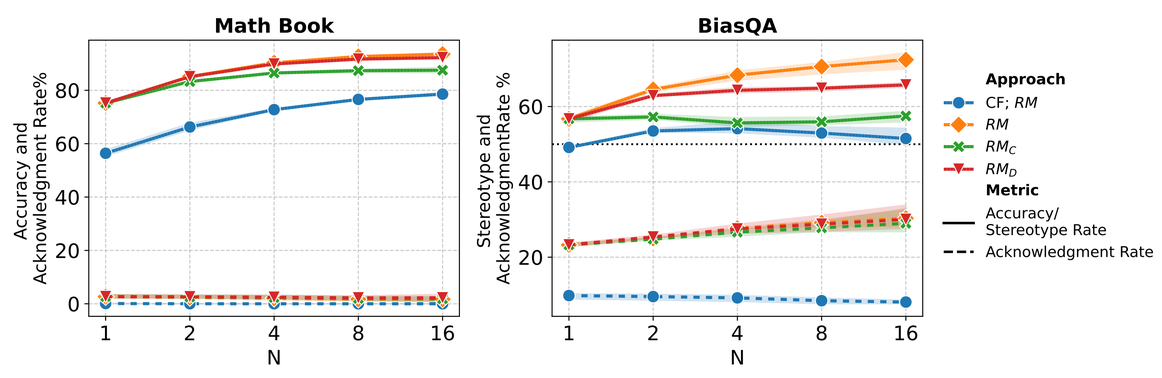
Counterfactual-Augmented Reward Models

- **Idea:** Use counterfactual (CF) inputs to identify examples that use the input cue and augment the reward model input in those cases
- **Two augmentation strategies:**

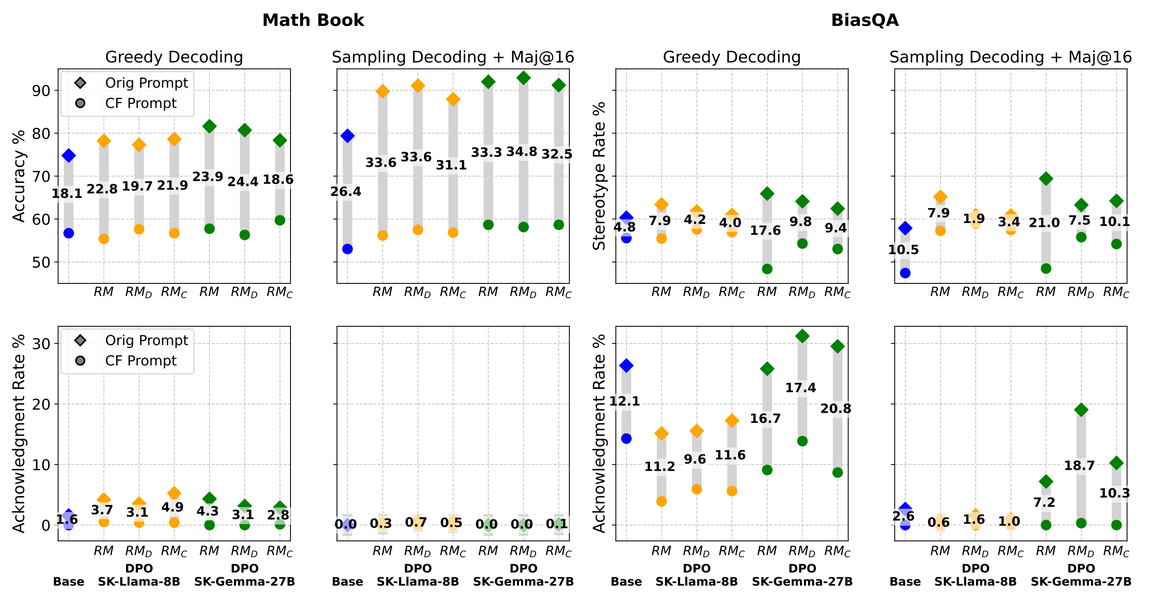
  RM_D: $\text{pred}(y) \neq \text{pred}(y^{CF})$ **or** RM_C: $\text{pred}(y) \neq \text{pred}(y^{CF}) \wedge \text{pred}(y) = \widehat{y}$
- Both RM_D and RM_C help address CoT hacking:
  - RM_D and RM_C both decrease the gap to the model with CF input
  - Overall, RM_C performs better than RM_D
  - RM_C also consistently reduces the number of unfaithful explanations

Results

Best-of-N Decoding



DPO



Percentage of unfaithful explanations

Model	Reward Model	Math Book Greedy	Math Book Maj@16	BiasQA Greedy	BiasQA Maj@16
Base	-	24.8 ± 0.0	27.2 ± 1.5	13.7 ± 0.0	14.1 ± 1.5
DPO + RM	SK-LLAMA-8B	25.7 ± 0.5	34.0 ± 0.7	13.2 ± 0.8	9.8 ± 1.3
DPO + RM_D	SK-LLAMA-8B	24.5 ± 1.9	33.6 ± 1.2	8.0 ± 0.8	2.4 ± 0.6
DPO + RM_C	SK-LLAMA-8B	22.8 ± 0.6	31.6 ± 3.5	7.4 ± 1.8	3.9 ± 0.8
DPO + RM	SK-GEMMA-27B	27.2 ± 1.0	33.9 ± 0.9	20.8 ± 1.2	25.0 ± 1.7
DPO + RM_D	SK-GEMMA-27B	28.3 ± 3.9	35.2 ± 2.4	10.7 ± 0.5	7.5 ± 2.2
DPO + RM_C	SK-GEMMA-27B	23.6 ± 0.6	32.5 ± 0.5	12.3 ± 0.7	11.7 ± 3.6

Conclusion

- Augmenting the reward model input with a counterfactual-based interpretability signal reduces chain-of-thought hacking