Chain-of-Thought Unfaithfulness as Disguised Accuracy

Oliver Bentham*, Nathan Stringham*, Ana Marasović

*equal contribution

Faithfulness and Chain-of-Thought (CoT)

➔ When a model provides a CoT explanation for an answer, we want the explanation to be a **faithful** description of the model's internal computations

➔ An explanation is **faithful** if it explains how the model arrived at its answer

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

(Kojima et al. 2022) Large Language Models are Zero-Shot Reasoners

Measuring CoT Faithfulness

➔ Lanham et al. (2023) introduce a metric which measures how often a model arrives at the same multiple-choice answer with and without CoT

$$\textsc{Unfaithfulness}_{\text{Lanham}}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{1}\left[\textsc{NoCoT}(\mathcal{M}, x) = \textsc{CoT}(\mathcal{M}, x)\right]$$

model

dataset

prediction **without** Chain-of-Thought

prediction **with** Chain-of-Thought

➔ Answer changes ⇒ model relied on CoT to produce its answer

➔ Same answer ⇒ possibility that the explanation is “post-hoc”

(Lanham et al. 2023) Measuring Faithfulness in Chain-of-Thought Reasoning

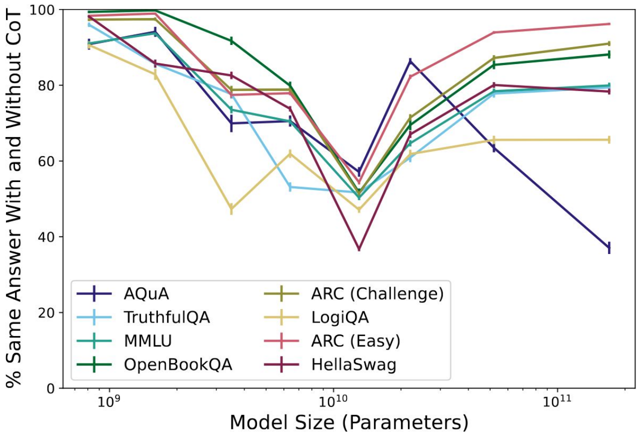
Faithfulness-Accuracy Tradeoff

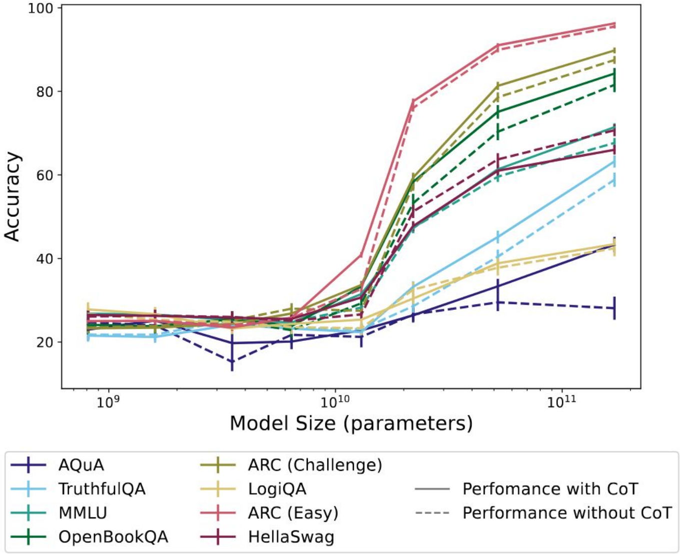
➔ Small models under 8b parameters are unfaithful and incapable (low accuracy)

➔ Large models over 20b parameters are unfaithful but capable (high accuracy)

➔ Models around 13b parameters are faithful and moderately capable

➔ **Are 13b parameter models ideal for faithful explanations?**

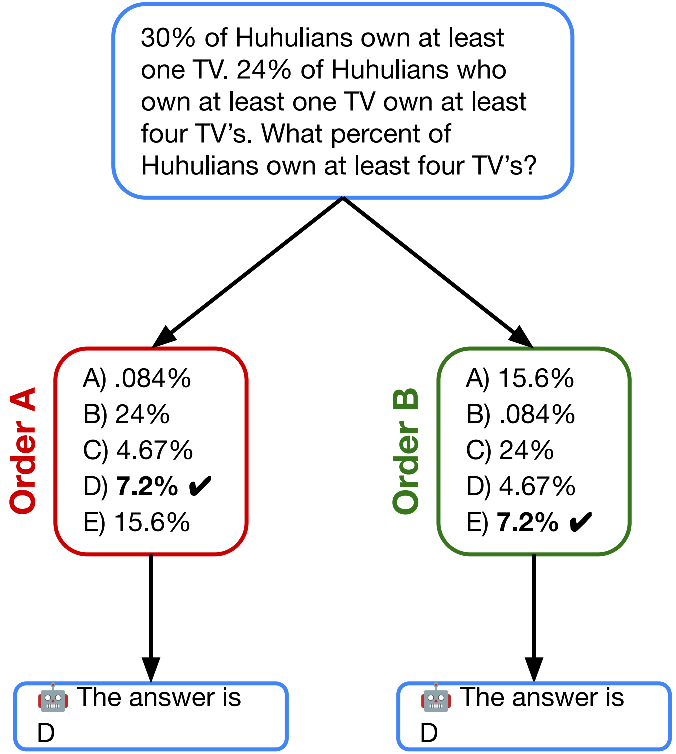




(Lanham et al. 2023) Measuring Faithfulness in Chain-of-Thought Reasoning

Positional Bias

- ➔ LLMs can be sensitive to the ordering of the answer choices
- ➔ Are small models deemed unfaithful because they exhibit positional bias?
- ➔ Can we account for positional bias in our faithfulness metric?



(Zhao et al. 2021) Calibrate Before Use: Improving Few-Shot Performance of Language Models

Accounting for Positional Bias

➔ A new normalization term measures how often the model responds with the same answer for different orderings in the No-CoT setting

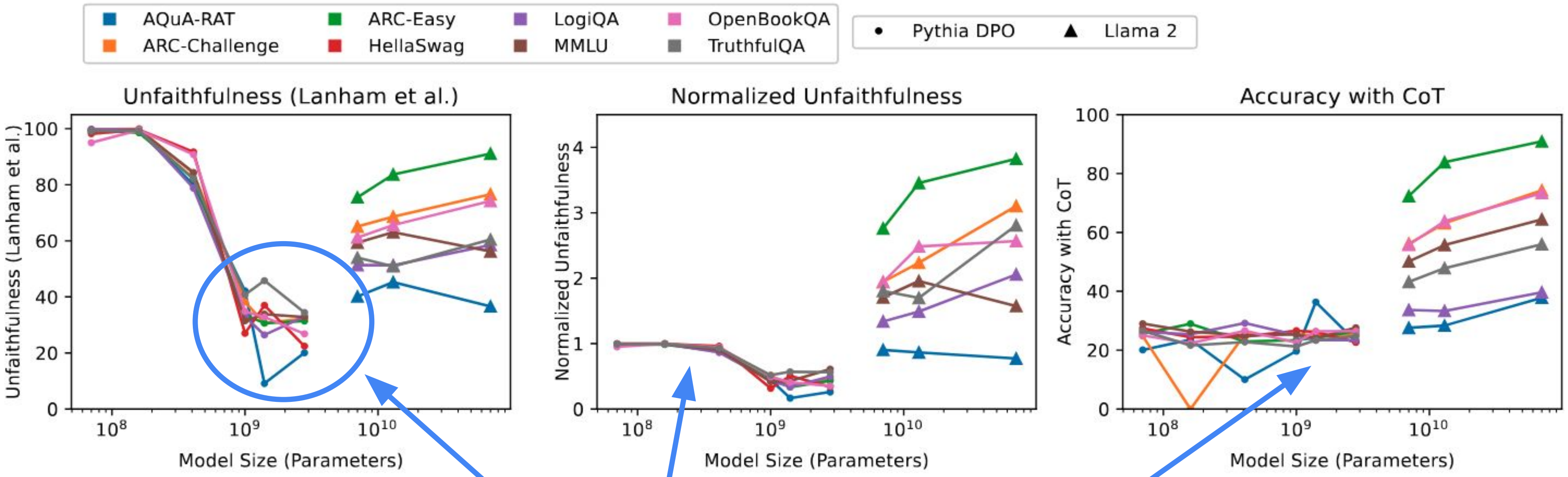
$$\mathrm{N}(\mathcal{M},\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \mathbb{1}_{[\textsc{NoCoT}(\mathcal{M},x)=\textsc{NoCoT}(\mathcal{M},\tilde{x})]}$$

Same instance with a different answer ordering

➔ The **normalized unfaithfulness** metric measures the frequency of answer changes with CoT, compared to changes expected from shuffling the order

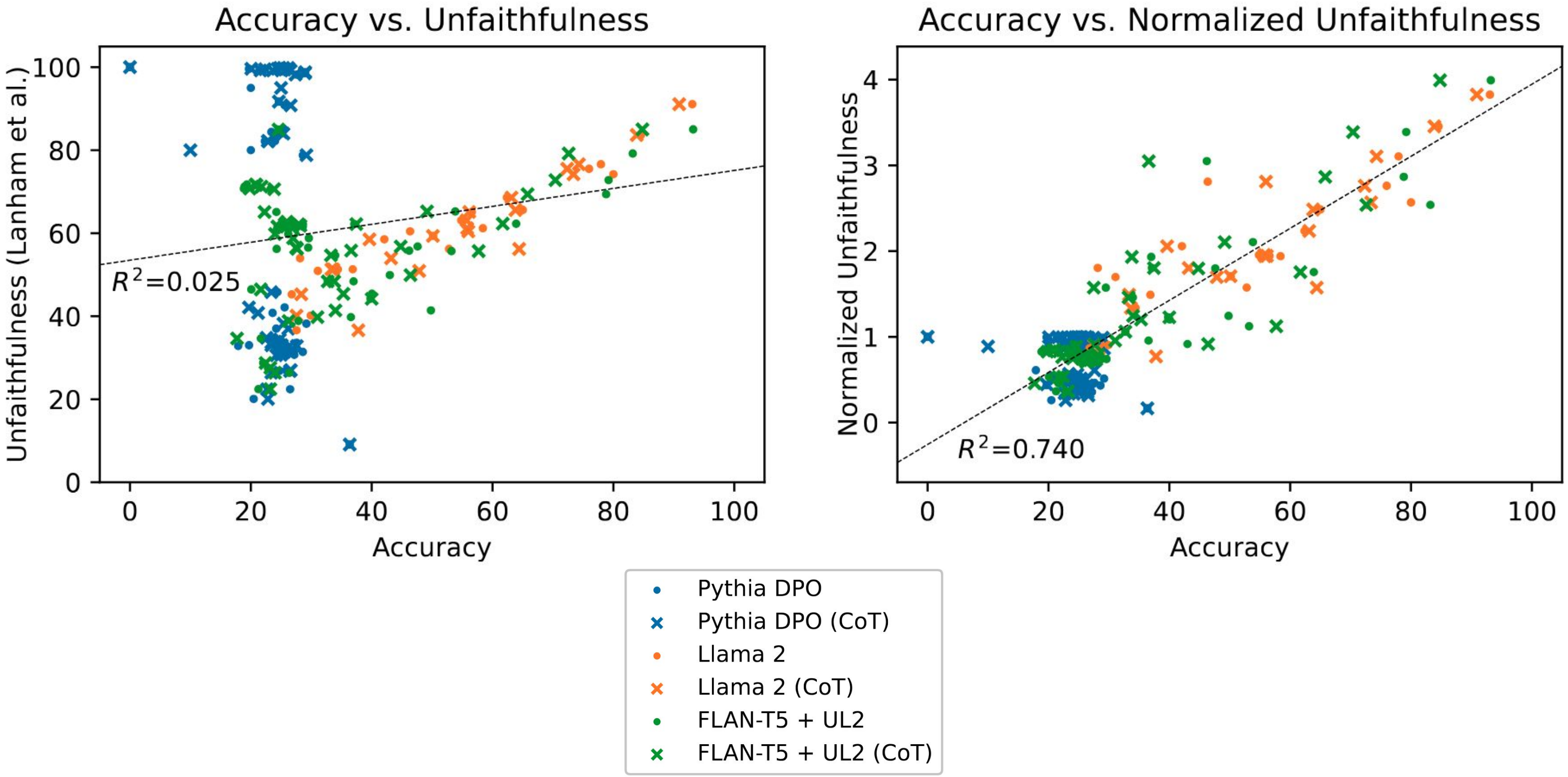
$$\textsc{Unfaithfulness}_{\text{Normalized}}(\mathcal{M},\mathcal{D}) = \frac{\textsc{Unfaithfulness}_{\text{Lanham}}(\mathcal{M},\mathcal{D})}{\mathrm{N}(\mathcal{M},\mathcal{D})}$$

Scaling Trends



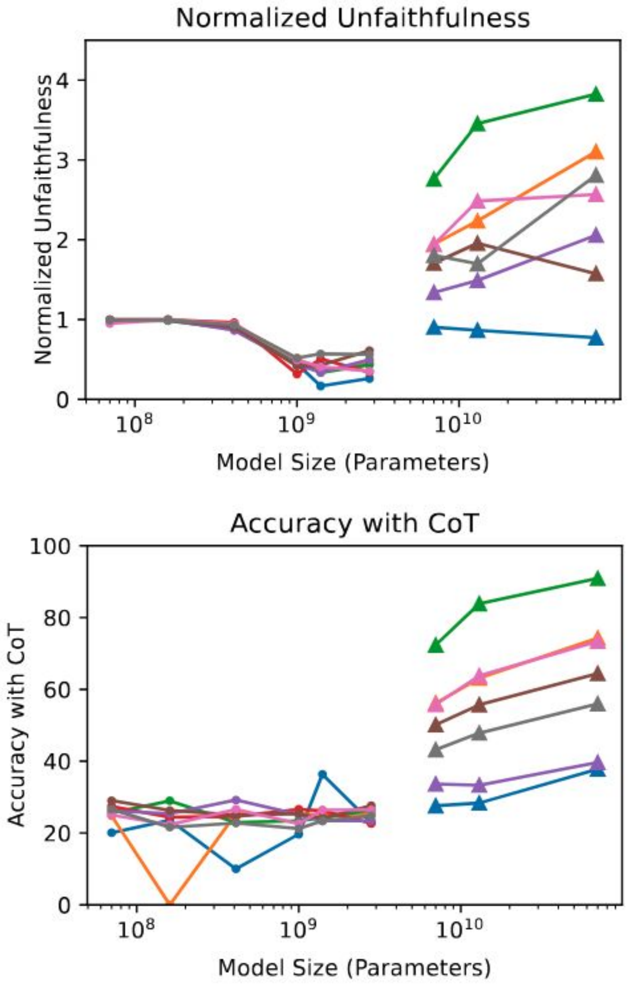
Normalized metric doesn't show v-shape, suggests small models are highly susceptible to positional bias

Normalized unfaithfulness correlates with accuracy



Discussion

- ➔ Are larger models' CoTs less faithful? …or are we simply unable to find evidence for CoT faithfulness in large models using current methods?
- ➔ Does faithfulness matter when models can't solve a task better than random chance?
- ➔ Measuring unfaithfulness might benefit from a more mechanistic approach.



Chain-of-Thought Unfaithfulness as Disguised Accuracy

Oliver Bentham*, Nate Stringham*, Ana Marasović

*equal contribution