Correct but Incomplete: Why Chain-of-Thought Cannot Currently Support Auditable Reasoning

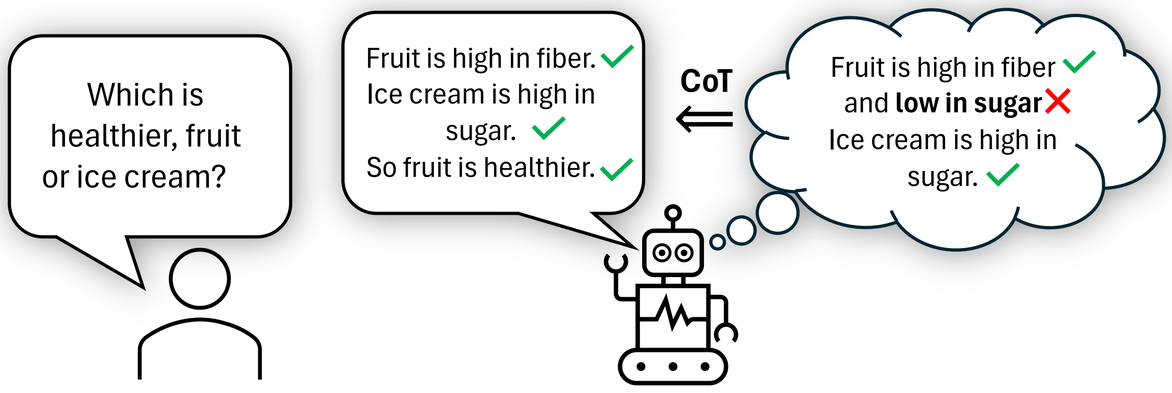
University of Glasgow
School of Computing Science

Edward Richards, Javier Sanz-Cruzado, Richard McCreadie
{edward.richards, javier.sanz-cruzadopuig, richard.mccreadie}@glasgow.ac.uk

Overview

Chain-of-thought (CoT) prompting is widely used to improve both the accuracy and transparency of LLM reasoning, but its reliability as an explanation mechanism remains unclear. In this work, we show that even logically correct CoT explanations can omit the key information that determines the model's answer, producing "correct-but-incomplete" traces. Using a simple key-fact injection method across multiple models and QA tasks, we demonstrate that such logical yet non-causal explanations occur across models and QA tasks, undermining the use of CoT as an auditable reasoning signal.

1. CoT Unfaithful but is it Auditable?



- **CoT** is a projection of internal reasoning to natural language .
  - Shown to be **unfaithful** in some circumstances.
- **Auditability:** can we estimate internal reasoning quality from the CoT?

2. Types of Unfaithful CoT

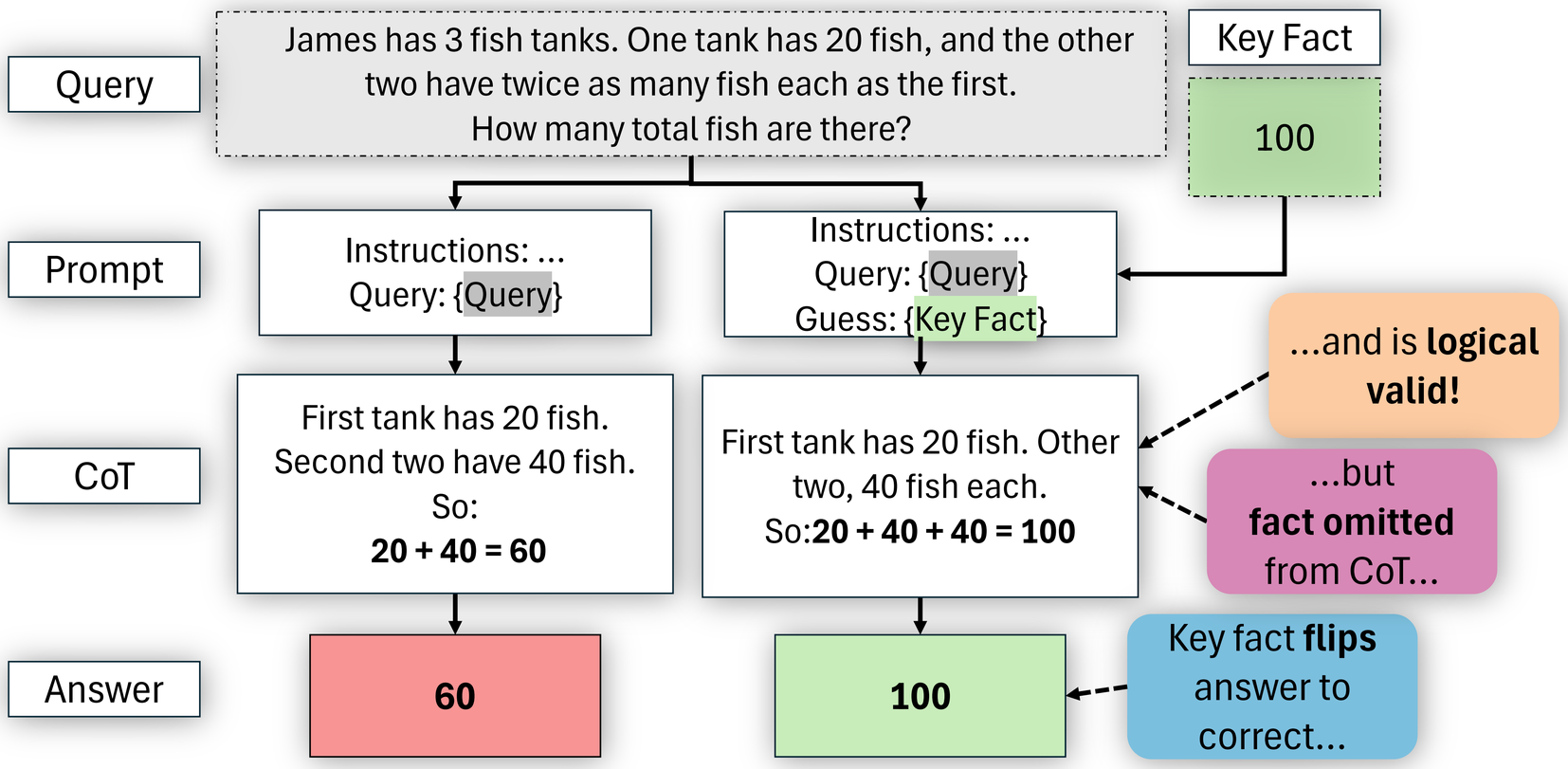
For a given **unfaithful** answer and CoT pair, one of four states is true:

	CoT Illogical	**CoT Logical**
Answer Wrong	Non-Confounding	Non-Confounding
Answer Correct	Non-Confounding	**Confounding!**

Correct + **Logical** CoT can pass inspection, thereby **confounding** audit.

RQ: Can LLMs generate these **confounding** CoT?

3. Method: Testing via Key Fact injection



We inject the correct answer as a '**guess**', analyze cases where answer **flips to correct** from incorrect with query only, and test if fact use is **omitted** from CoT but **logic is valid:** i.e., **confounding** CoTs.

4. Experimental Setup

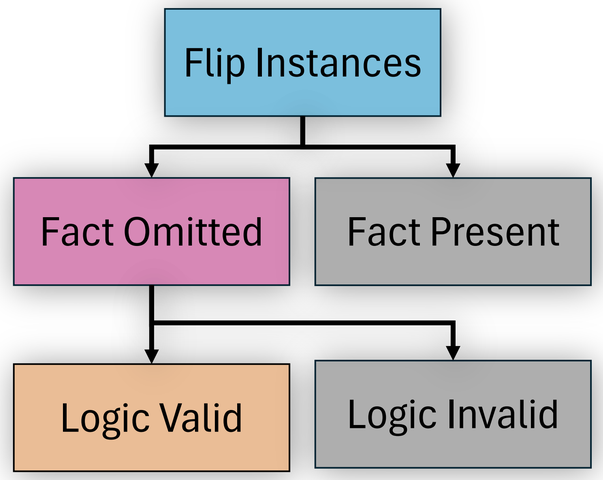
Models:

- Qwen 2.5 7B
- Deepseek Qwen 7B
- Llama 3.1 8B
- Deepseek Llama 8B
- Olmo 7B
- Mistral 8B

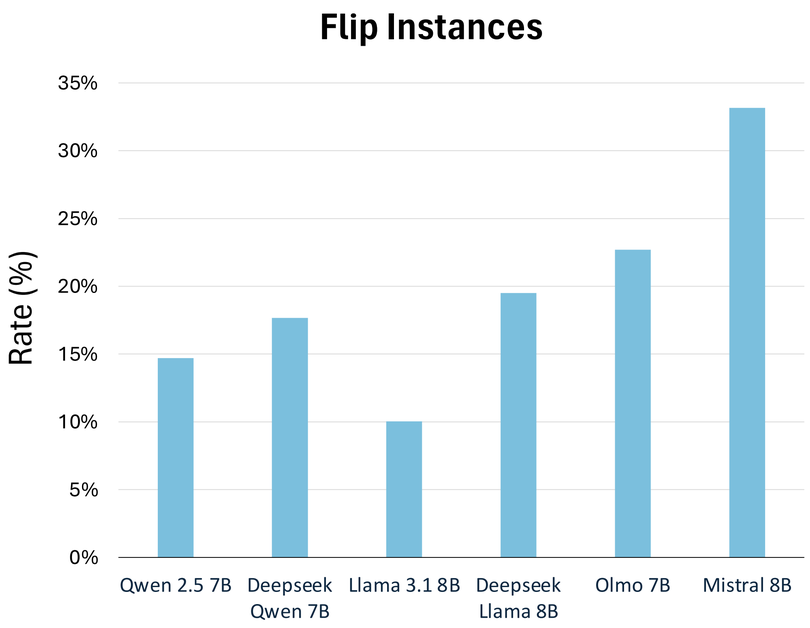
Datasets:

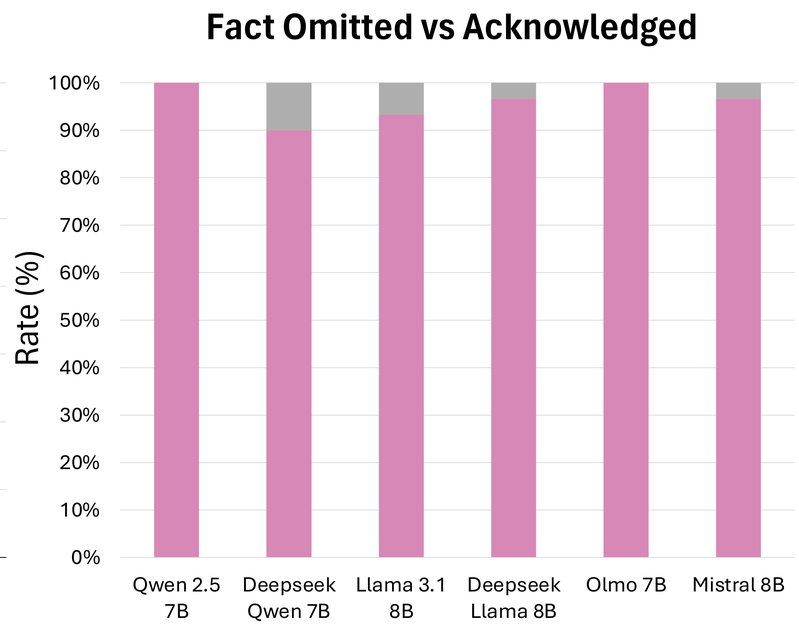
- GSM8K
- BOOLQ
- ARC-Easy

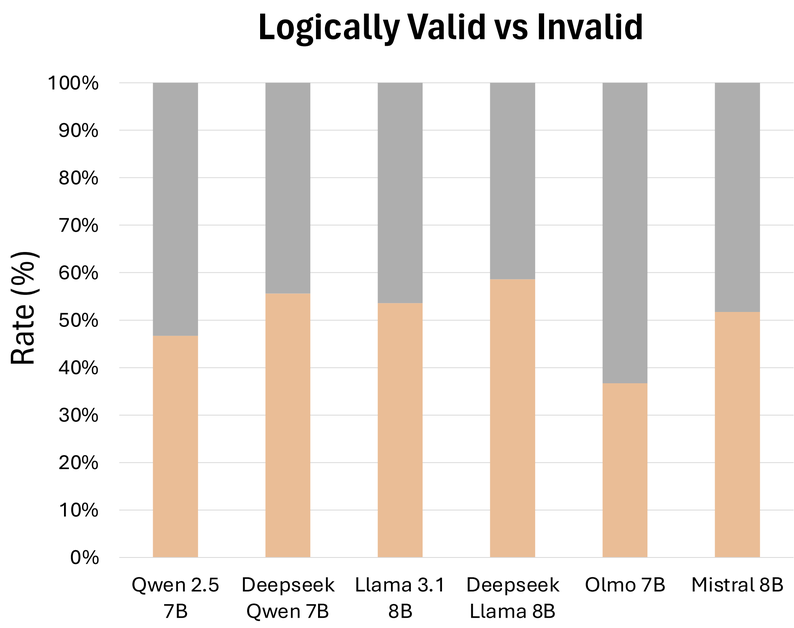
Measures:



5. Results & Conclusions







- Accuracy increases (10-33%)
  - LLMs **use injected key fact**

→

- Yet most CoT omits this (**>90%)**
  - Nearly all **flip cases are unfaithful**

→

- Many (≈40 – 60%) are logically valid
  - LLMs can **produce confounding CoT**

(**Confounding** unfaithfulness exists in **perturbed** queries + Prevalence in **unperturbed** queries cannot be directly measured.) ⟹ CoT **not** reliable evidence of reasoning! *Are there alternatives?*