Probabilistic Soundness Guarantees in LLM Reasoning Chains

Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin,
Chris Callison-Burch, Eric Wong
University of Pennsylvania

Penn Engineering

ICML International Conference On Machine Learning

LLMs often make reasoning errors

Context

Base Claim 1: The denominator of a fraction is 7 less than 3 times the numerator.
Base Claim 2: If the fraction is equivalent to 2/5, what is the numerator?

Correct Reasoning Chain

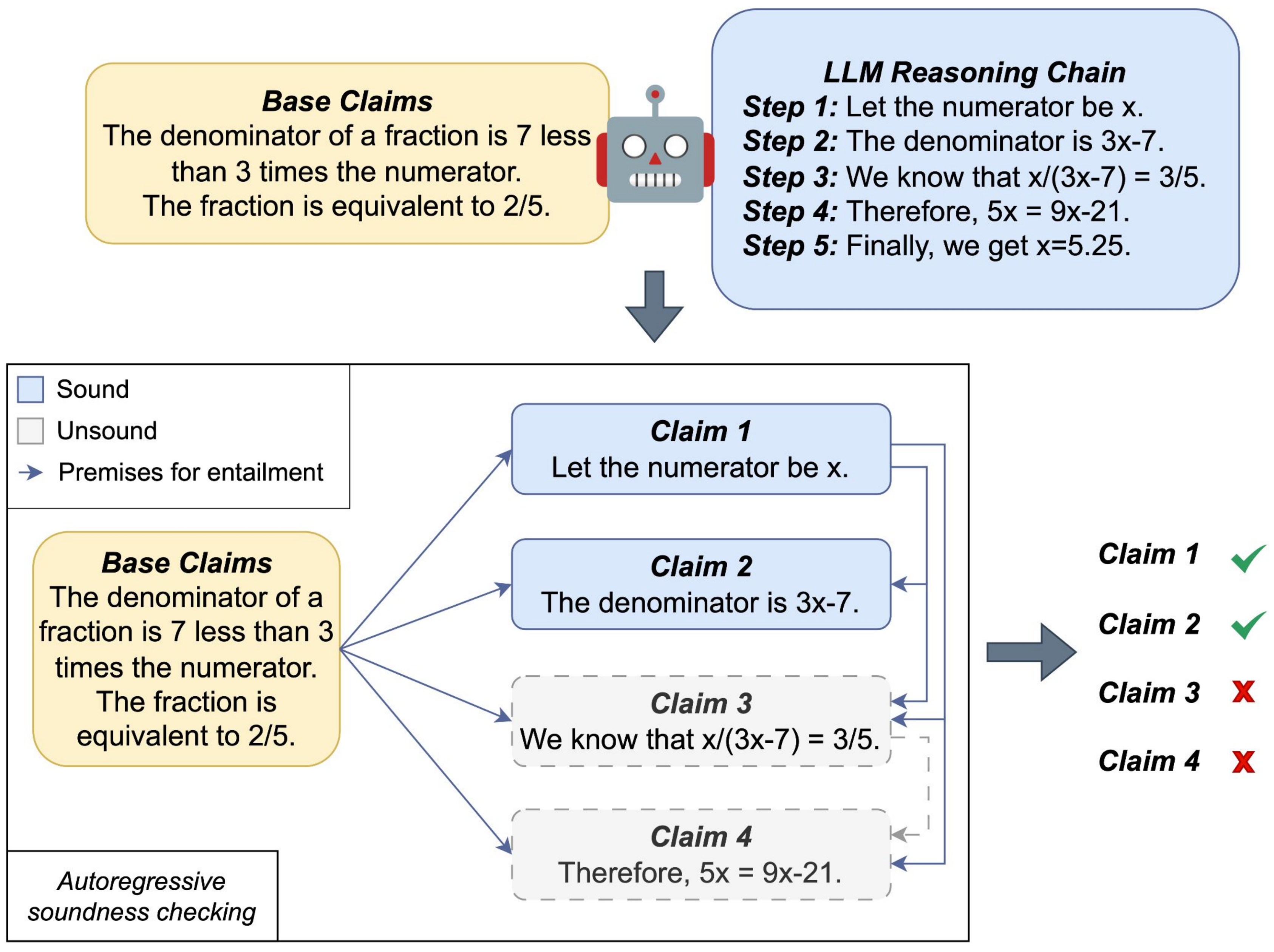
Step 1: Let the numerator be x.
Step 2: The denominator is 3x-7.
Step 3: We know that x/(3x-7) = 2/5.
Step 4: Therefore, 5x = 6x-14.
Step 5: Finally, we get x = 14. (Correct)

Unsound Steps

Step 1: Let the numerator be x.
Step 2: The denominator is 3x-7.
Step 3: We know that x/(3x-7) = 3/5
Step 4: Therefore, 5x = 9x-21
Step 5: Finally, we get x = 6. (Incorrect)

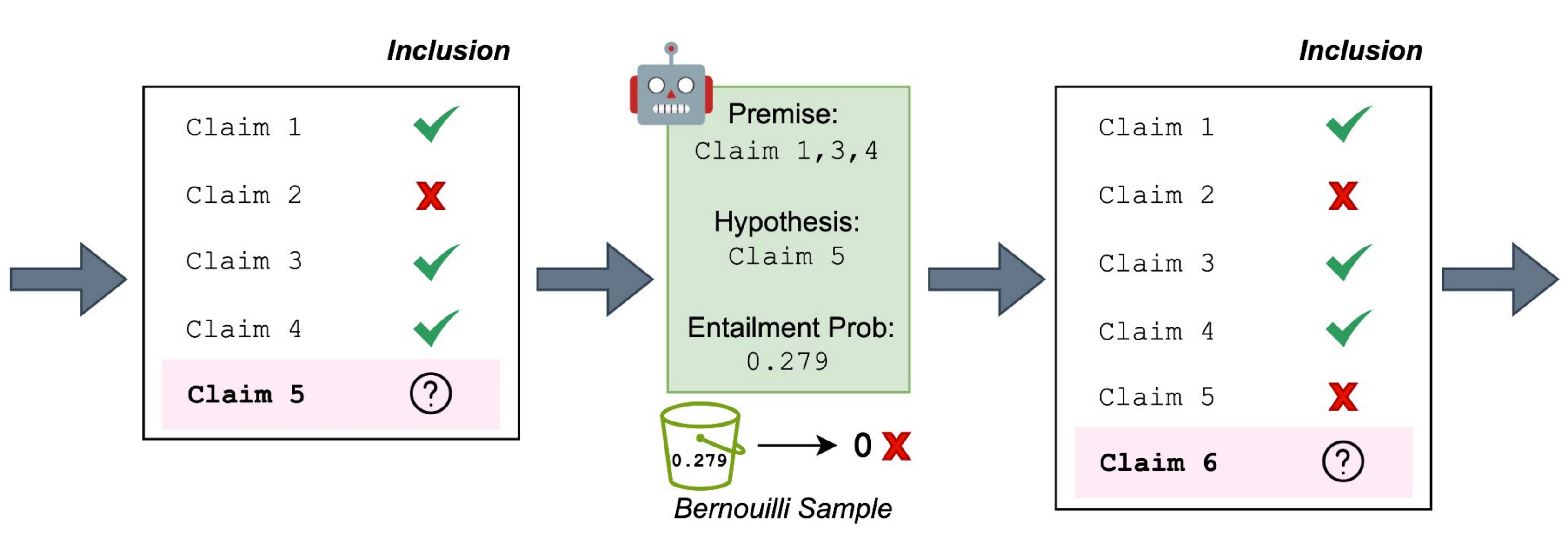
Existing methods struggle to detect ungrounded statements, propagated errors, and invalid derivations.

ARES: Structured error detection with logic



The entailment model autoregressively checks each claim with respect to the previous claims verified to be sound.

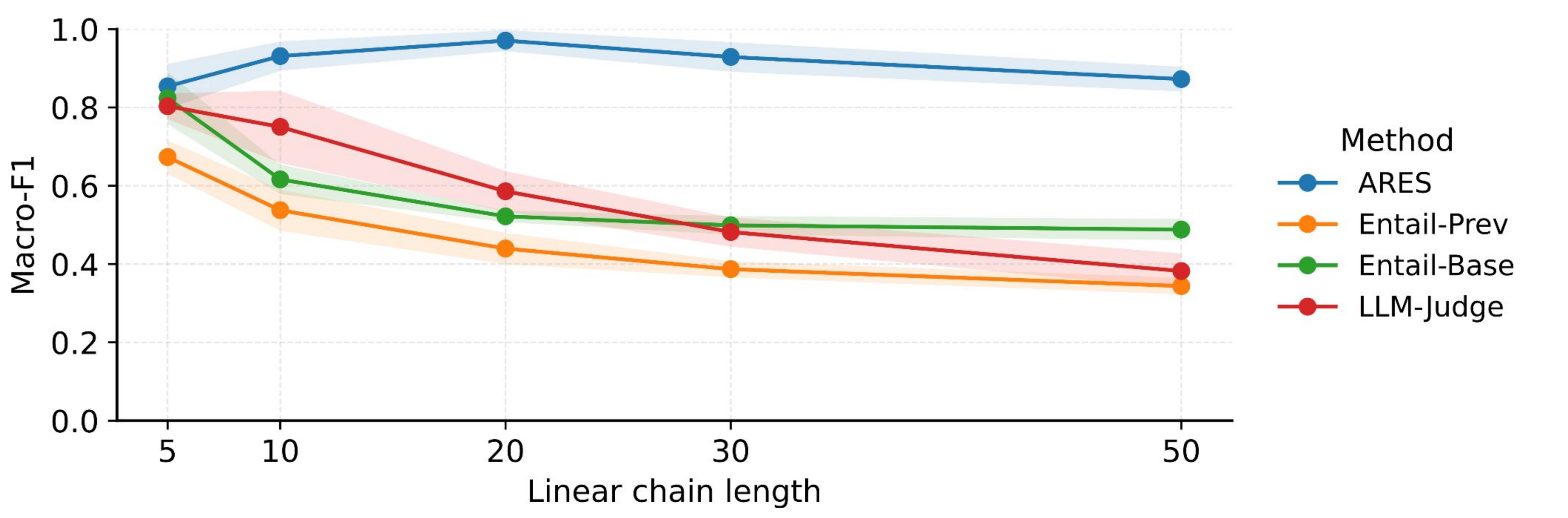
Statistical guarantees on reasoning



*Entailment models are **probabilistic**: each step's soundness estimated by probabilistically including previous claims by their soundness rates.*

Theorem. With $N \geq \log(2m/\delta)/(2\varepsilon^2)$ samples, *the soundness rate for m claims is estimated to* ±ε error with 1 - δ confidence.

ARES excels on long synthetic reasoning



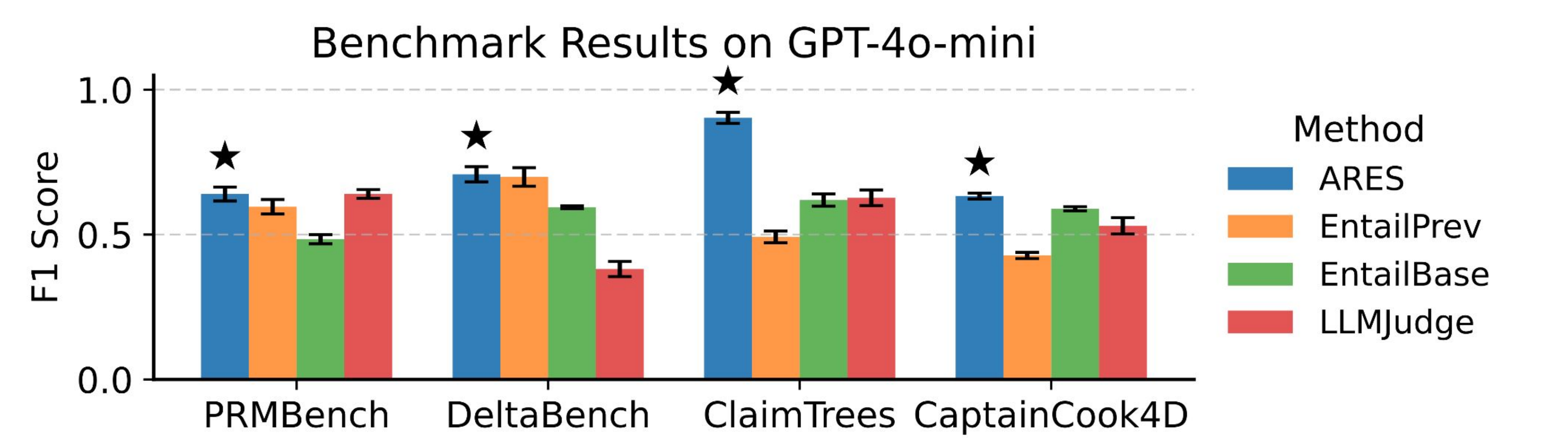
ARES maintains high Macro-F1 on ClaimTrees, even for long chains.

ClaimTrees: A synthetic reasoning benchmark

Reasoning Chain	ARES	Entail-Prev	Entail-Base	LLM-Judge
Base 1: Rule: H3 -> AZ	-	-	-	-
Base 2: Fact: I have D8	-	-	-	-
...	-	-	-	-
Base 9: Rule: DG -> G8	-	-	-	-
Claim 1: I have D8, I use rule (D8 -> U8) to derive U8	0.96	1.00	1.00	1.00
...	...	...	...	...
Claim 6: I have H3, I use rule (H3 -> AZ) to derive AZ	0.90	1.00	1.00	1.00
Claim 7: I have AZ, I use rule (AZ -> SG) to derive SG	0.00	0.00	0.00	0.20
Claim 8: I have SG, I use rule (SG -> C6) to derive C6	0.09	1.00	1.00	1.00

Only ARES detects the propagated error: The non-existent rule (AZ -> SG) cannot be used!

ARES also wins on real benchmarks



Why is ARES so effective?

Method	Robust	Causal	Sufficient
ARES (ours)	✓	✓	✓
Entail-Prev	✗	✓	✓
Entail-Base	✓	✓	✗
LLM-Judge	✗	✗	✓

Robust: Previous errors do not adversely affect current step.
Causal: Downstream steps do not affect current step.
Sufficient: All relevant claims included as premise for detection.

BRACHIO LAB