

Attribute units in a given LLM output to prior units, causally 🍊

Local Causal Attribution of Chain-of-Thought Reasoning

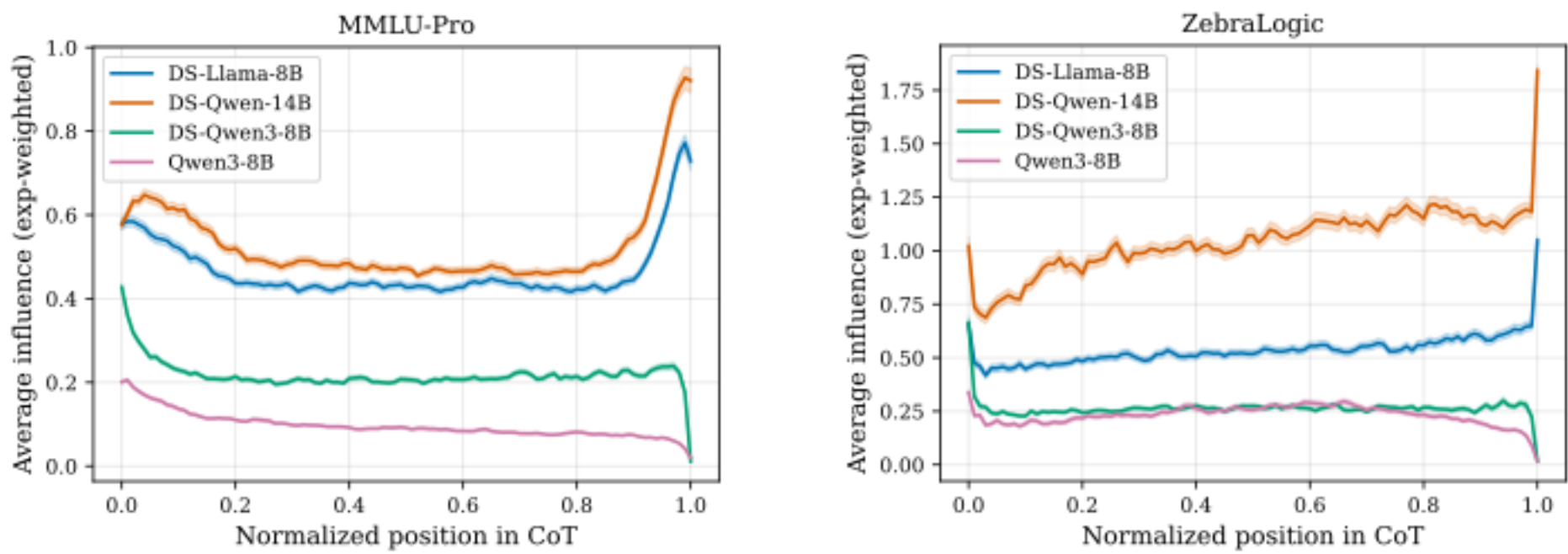
Goal: Investigate causal structure within a single, given CoT (“locally”) in a black-box manner

- Not distributions over CoT steps/trajectories, not concerned with faithfulness of CoT text

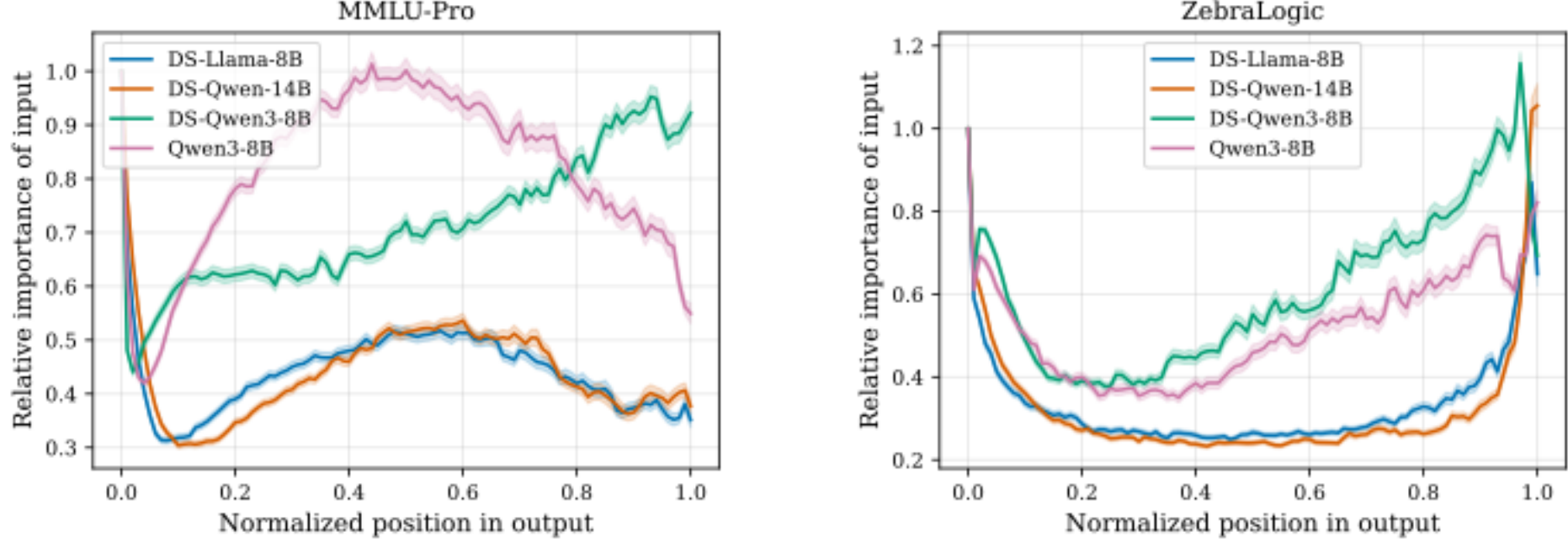
Our **AttriCoT** method attributes output units (columns) to prior units (rows) by estimating effects on log probability.

Dataset-level analysis of AttriCoT’s attribution scores shows differences between reasoning models.

	[CoT_1] First, I recognize t...	[CoT_2] Next, the problem st...	[CoT_3] Finally, to determin...	[Ans_1] To determine the tot...	[Ans_2] 1. **Blue Fiber:** ...	[Ans_3] 2. **White Fiber:** ...	[Ans_4] 3. **Total Bolts Nee...	[Ans_5] \boxed{3}
[In_1] A robe takes 2 bolts of blue fiber and half that much white fiber.	2.25	1.70	0.06	0.04	0.05	0.03	0.00	-0.01
[In_2] How many bolts in total does it take?	0.06	0.00	0.37	0.03	0.01	0.00	-0.02	-0.02
[CoT_1] First, I recognize that the robe requires 2 bolts of blue fiber.		1.56	0.00	0.03	0.06	0.03	0.01	0.02
[CoT_2] Next, the problem states that it requires half as much white fiber as blue fiber. To find the amount of white fiber needed, I calculate half of 2 bolts, which is 1 bolt.			1.31	0.02	0.01	-0.01	0.00	0.00
[CoT_3] Finally, to determine the total number of bolts required, I add the blue and white fibers together: 2 bolts + 1 bolt equals 3 bolts in total.				0.88	0.00	0.01	0.00	0.02
[Ans_1] To determine the total number of bolts needed for the robe, let's break down the requirements step by step.					2.60	-0.01	0.01	-0.04
[Ans_2] 1. **Blue Fiber:** - The robe requires **2 bolts** of blue fiber.						1.00	0.07	0.02
[Ans_3] 2. **White Fiber:** - The amount of white fiber needed is **half** of the blue fiber. - Calculation: $\frac{1}{2} \times 2 = 1$ bolt = **1 bolt** of white fiber.							1.25	0.24
[Ans_4] 3. **Total Bolts Needed:** - Add the bolts of blue and white fiber together. - Calculation: $(2 \text{ bolts (blue)}) + (1 \text{ bolt (white)}) = 3 \text{ bolts}$.								0.05
[Ans_5] \boxed{3}								

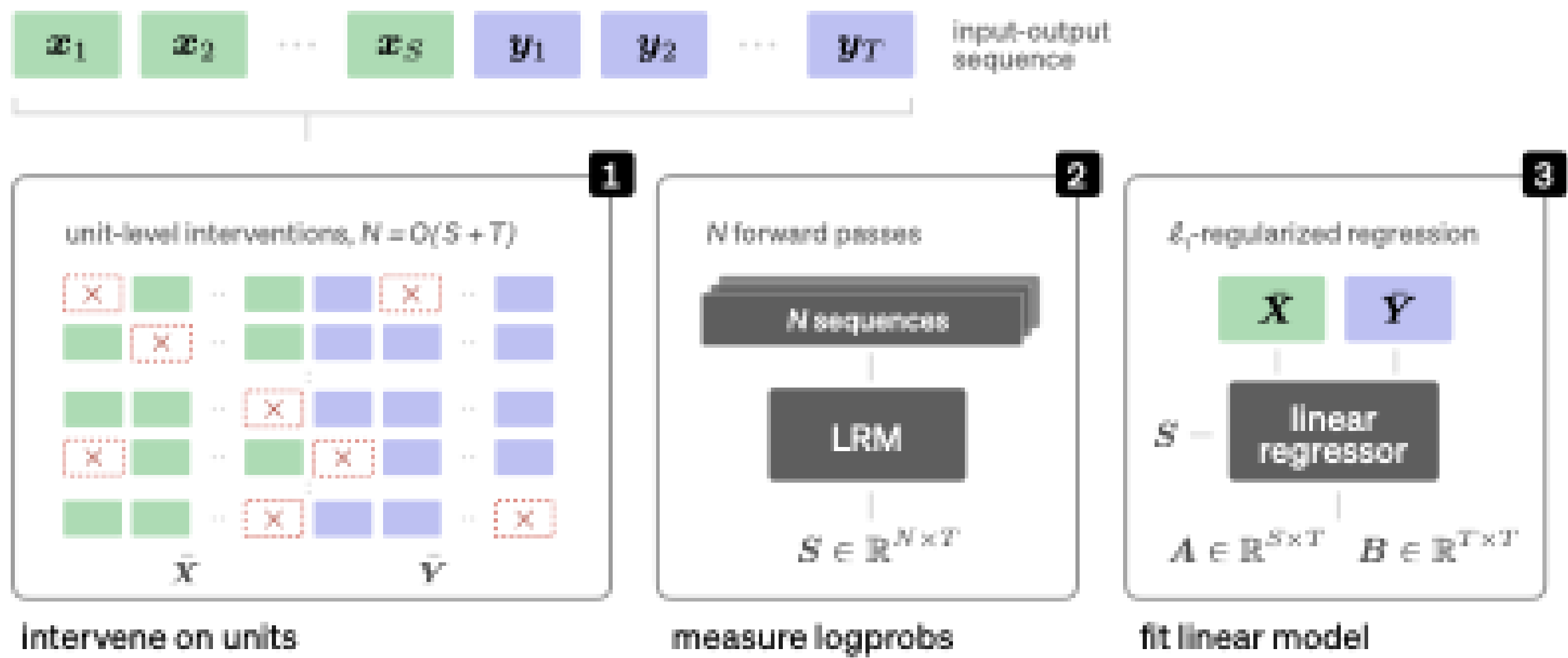


Middle CoT units have constant influence on later units. Distinctive behavior occurs at beginnings and ends.



Input units gradually decrease in importance, then increase again, suggesting that models refer back to the input.

AttriCoT Method



AttriCoT’s attributions are consistently more faithful to the model’s behavior than Thought Anchors [1] and prompting baselines (see paper).

[1] Bogdan et al. Thought anchors: Which LLM reasoning steps matter? *Mech Interp Workshop* 2025.

AttriCoT fits a linear structural causal model:

$$\log P(y_t|x,y_{<t}) = \sum_{s=1}^S \alpha_{s,t} \mathbb{I}(x_s) + \sum_{\tau=1}^{t-1} \beta_{\tau,t} \mathbb{I}(y_\tau) + \gamma_t + u_t, \quad t = 1, \dots, T$$

log probability of output unit presence/absence of prior units



PAPER

Dennis Wei, Yannis Belkhiter, Erik Miehl, Radu Marinescu

Mechanistic Interpretability Workshop, ICML 2026

