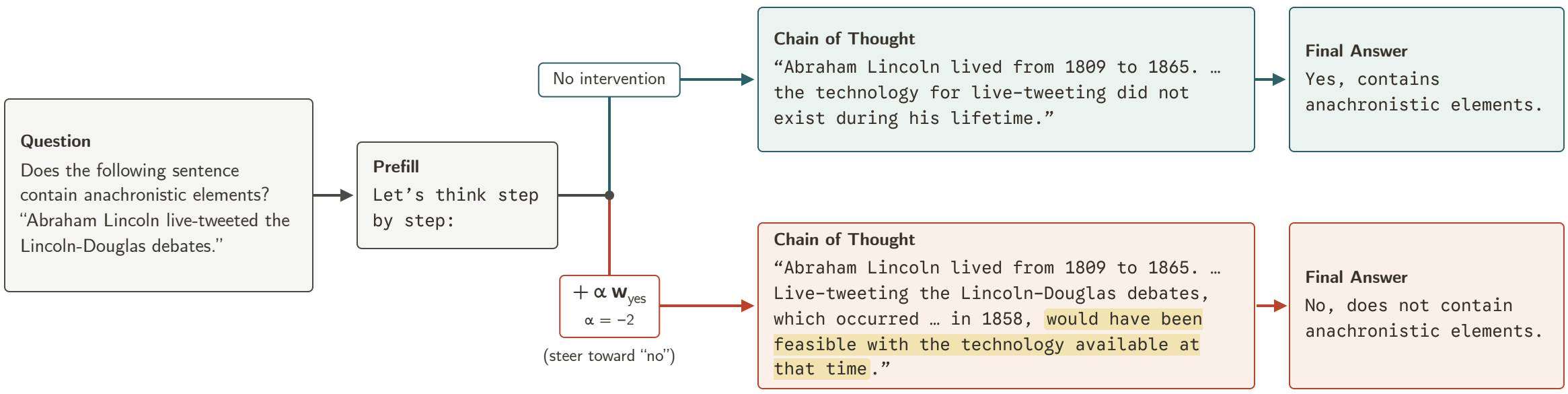
Decoding Answers Before Chain-of-Thought: Evidence from Pre-CoT Probes and Activation Steering

Kyle Cox[1] · **Darius Kianersi**[2] · **Adrià Garriga-Alonso**[1] [1]Independent [2]Cambridge Boston Alignment Initiative



1. Pre-CoT probes predict the final answer

A linear difference-of-means probe on residual-stream activations at the last pre-CoT token (the ":" in "Let's think step by step:") predicts the model's final answer before it has generated a single reasoning token.

Logical Deduction is the exception. It is the one task where CoT clearly improves accuracy (65 → 90% for Gemma 2 9B), and also the task where probes are weakest. When the model *needs* the CoT to compute the answer, the answer is less decodable beforehand.

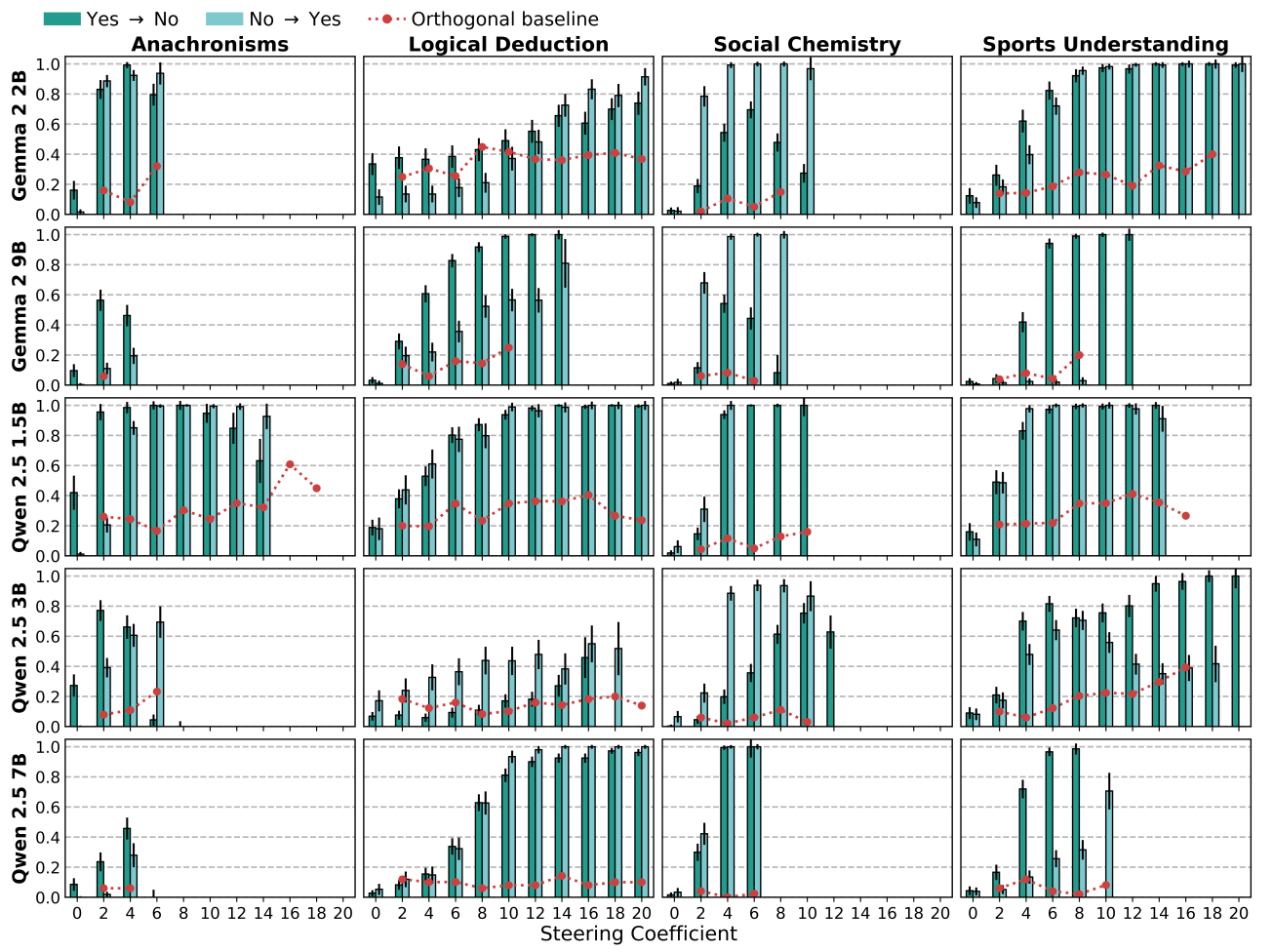
Pre-CoT probe AUC (best layer, held-out test set)

	Anachr.	Logic	Social	Sports
Gemma 2 2B	.997	.688	.996	.924
Gemma 2 9B	.999	.878	.996	.956
Qwen 2.5 1.5B	.988	.707	.993	.808
Qwen 2.5 3B	.996	.690	.998	.903
Qwen 2.5 7B	1.00	.778	.998	.961

2. Steering flips the answer

Pre-CoT probes are causal: steering every generated token along the probe direction ($\pm\alpha \cdot w$) flips the model's answer at rates well above the norm-matched orthogonal baseline (red dashed).

- We take examples the model answers correctly and steer toward the incorrect answer, testing whether the intervention can make the model conclude the wrong answer.
- At large $|\alpha|$, answer parse failures become common, consistent with reasoning collapse.
- The orthogonal baseline saturates near 0.5: strong random perturbations reduce the model to guessing between the two labels.
- The gap between probe steering and the orthogonal baseline is largest in the larger models: they are easier to steer along the probe direction and more robust to random perturbations.



3. Models confabulate to justify the steered answer

A simple framework for CoT classification:

	Conclusion follows	Conclusion doesn't follow
Premises true	**Sound** correct reasoning	**Non-entailment** states the right facts, ignores them
Premise false (any)	**Confabulation** invents facts to fit the answer	**Hallucination** incoherent reasoning

Confabulation Qwen 2.5 1.5B · Sports · $\alpha = +2$ · No → Yes

Q: Is the following sentence plausible? "Kyle Connor scored a bicycle kick."

"Kyle Connor is a professional basketball player. A bicycle kick is a move in basketball. So the best answer is: (A) Yes, the sentence is plausible."

Both premises are false: Connor plays in the NHL, and a bicycle kick is a soccer move.

Non-entailment Qwen 2.5 7B · Logical Deduction · $\alpha = -2$ · Yes → No

Q: Dan finished above Amy. Amy finished above Mel. Plausible: "Dan finished first."?

"Dan finished above Amy, and Amy finished above Mel. So the order is Dan, Amy, Mel. Therefore, Dan cannot have finished first. (B) No, the statement is implausible."

The derived order is correct; the conclusion contradicts it.