Model	Method	Forward Belief			Forward Action			Backward Belief		
		TB	FB	Both	TB	FB	Both	TB	FB	Both
Llama-2-7b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$44_{0.0}$	$44_{0.0}$	$44_{0.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$66_{2.0,11}$	$71_{1.0,31}$	$54_{2.0,0}$	$66_{2.0,11}$	$57_{2.0,12}$	$54_{2.0,2}$	$60_{2.0,11}$	$74_{1.0,31}$	$54_{2.0,2}$
Llama-2-7b-chat	No int.	56	56	55	69	55	37	56	56	55
	ITI	$58_{15.0}$	$58_{15.0}$	$57_{15.0}$	$69_{0.0}$	$55_{0.0}$	$37_{0.0}$	$58_{10.0}$	$60_{10.0}$	$57_{10.0}$
	CAA	$70_{1.0,11}$	$72_{1.5,10}$	$57_{1.0,1}$	$69_{0.0,0}$	$67_{1.5,11}$	$53_{1.5,12}$	$66_{1.0,11}$	$84_{1.5,10}$	$57_{1.0,0}$
Llama-2-13b	No int.	52	44	35	59	50	37	46	49	33
	ITI	$52_{0.0}$	$45_{15.0}$	$35_{0.0}$	$64_{15.0}$	$61_{20.0}$	$46_{20.0}$	$48_{20.0}$	$59_{20.0}$	$42_{20.0}$
	CAA	$85_{2.0,12}$	$88_{2.0,14}$	$66_{2.0,12}$	$71_{1.5,10}$	$69_{2.0,13}$	$55_{1.0,39}$	$75_{2.0,10}$	$92_{2.0,13}$	$59_{1.5,12}$
Llama-2-13b-chat	No int.	84	56	47	78	51	38	72	48	31
	ITI	$84_{0.0}$	$65_{15.0}$	$59_{15.0}$	$78_{0.0}$	$58_{15.0}$	$47_{15.0}$	$72_{0.0}$	$60_{15.0}$	$48_{15.0}$
	CAA	$97_{1.0,12}$	$94_{1.0,12}$	$91_{1.0,12}$	$80_{1.5,11}$	$71_{1.0,13}$	$54_{1.5,13}$	$97_{1.5,10}$	$94_{1.5,12}$	$87_{1.5,12}$
Llama-2-70b	No int.	90	87	78	93	52	48	73	53	32
	ITI	$90_{0.0}$	$90_{20.0}$	$78_{0.0}$	$94_{15.0}$	$55_{20.0}$	$50_{15.0}$	$77_{10.0}$	$58_{15.0}$	$37_{10.0}$
	CAA	$99_{2.0,16}$	$97_{1.5,19}$	$95_{1.5,18}$	$94_{1.5,2}$	$80_{2.0,19}$	$73_{1.5,18}$	$94_{2.0,18}$	$92_{2.0,19}$	$83_{1.5,19}$
Llama-2-70b-chat	No int.	69	75	56	86	56	52	63	59	52
	ITI	$69_{0.0}$	$76_{10.0}$	$59_{10.0}$	$86_{0.0}$	$56_{0.0}$	$52_{0.0}$	$63_{0.0}$	$60_{10.0}$	$54_{10.0}$
	CAA	$92_{1.5,18}$	$97_{1.5,25}$	$89_{1.5,18}$	$87_{1.5,17}$	$75_{1.0,19}$	$60_{1.0,19}$	$88_{1.5,18}$	$92_{1.0,19}$	$80_{1.5,18}$
Pythia-70m	No int.	41	41	37	46	45	41	44	41	37
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$62_{1.0,2}$	$56_{1.0,1}$	$54_{1.5,1}$	$59_{1.0,2}$	$60_{1.0,3}$	$58_{1.0,2}$	$63_{1.0,2}$	$56_{1.0,2}$	$54_{1.5,1}$
Pythia-410m	No int.	48	45	45	44	44	44	44	47	44
	ITI	$55_{20.0}$	$62_{20.0}$	$52_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$60_{20.0}$	$63_{20.0}$	$56_{20.0}$
	CAA	$67_{2.0,4}$	$64_{2.0,4}$	$61_{2.0,0}$	$56_{2.0,6}$	$63_{1.5,12}$	$56_{2.0,6}$	$69_{2.0,4}$	$63_{2.0,0}$	$60_{2.0,0}$
Pythia-1b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$59_{2.0,8}$	$62_{2.0,5}$	$54_{2.0,0}$	$57_{2.0,4}$	$59_{2.0,10}$	$56_{2.0,4}$	$57_{2.0,3}$	$60_{2.0,5}$	$54_{2.0,0}$
Pythia-6.9b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$45_{20.0}$	$54_{20.0}$	$44_{0.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$56_{1.5,12}$	$71_{1.5,9}$	$55_{2.0,23}$	$55_{2.0,4}$	$63_{1.5,11}$	$55_{2.0,4}$	$55_{2.0,23}$	$71_{1.5,9}$	$55_{2.0,23}$
Pythia-6.9b-chat	No int.	55	54	28	36	64	20	44	67	30
	ITI	$57_{15.0}$	$54_{0.0}$	$28_{0.0}$	$44_{15.0}$	$71_{15.0}$	$32_{15.0}$	$44_{0.0}$	$67_{0.0}$	$30_{0.0}$
	CAA	$68_{1.5,15}$	$65_{1.5,12}$	$57_{1.5,11}$	$54_{1.5,10}$	$75_{1.5,5}$	$48_{1.5,10}$	$58_{1.5,15}$	$67_{0.0,0}$	$54_{1.5,10}$
Pythia-12b	No int.	44	44	44	44	44	44	44	44	44
	ITI	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$	$54_{20.0}$
	CAA	$54_{2.0,0}$	$64_{2.0,9}$	$54_{2.0,0}$	$60_{2.0,11}$	$58_{2.0,11}$	$55_{2.0,12}$	$54_{2.0,0}$	$67_{2.0,10}$	$54_{2.0,0}$

Table 3: Activation intervention: comparison between ITI (Li et al., 2023b) and CAA (Rimsky et al., 2023). For ITI, the subscript indicates the value of the coefficient α_{ITI} used: $\text{Acc}_{\alpha_{\text{ITI}}}$. For CAA, the subscript indicates first the value of the coefficient α used and second the layer l at which intervention takes place: $\text{Acc}_{\alpha_{\text{CAA}},l}$.

Model	Method	Control	CAA Parameters
Llama-2-7b	No int.	44	
	CAA	66_{+22}	2.0, 11
Llama-2-7b-chat	No int.	56	
	CAA	70_{+14}	1.0, 11
Llama-2-13b	No int.	52	
	CAA	85_{+33}	2.0, 12
Llama-2-13b-chat	No int.	84	
	CAA	97_{+13}	1.0, 12
Llama-2-70b	No int.	90	
	CAA	99_{+9}	2.0, 16
Llama-2-70b-chat	No int.	69	
	CAA	92_{+23}	1.5, 18
Pythia-70m	No int.	41	
	CAA	62_{+21}	1.0, 2
Pythia-410m	No int.	48	
	CAA	67_{+19}	2.0, 4
Pythia-1b	No int.	44	
	CAA	59_{+15}	2.0, 8
Pythia-6.9b	No int.	44	
	CAA	56_{+12}	1.5, 12
Pythia-6.9b-chat	No int.	55	
	CAA	68_{+13}	1.5, 15
Pythia-12b	No int.	44	
	CAA	54_{+10}	2.0, 0

Table 4: Results for CAA (Rimsky et al., 2023) on the Forward Belief True Control task in BigToM (Gandhi et al., 2023). Numbers indicate accuracy scores, with differences (CAA − No int.) as subscripts.

Example 1: Forward Belief

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk. Noor does not see her coworker swapping the milk.
Does Noor believe the milk pitcher contains oat milk or almond milk?
a) Noor believes the milk pitcher contains oat milk.
b) Noor believes the milk pitcher contains almond milk.

Example 2: Forward Action

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk. Noor does not see her coworker swapping the milk.
What will Noor do?
a) Noor will make the cappuccino using the milk in the pitcher.
b) Noor will open the fridge once again to take out the oat milk and replace the almond milk with oat milk.

Example 3: Backward Belief

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor opens the fridge again and reaches for the oat milk. Noor makes the cappuccino using the milk in the pitcher.
Does Noor believe the milk pitcher contains oat milk or almond milk?
a) Noor believes the milk pitcher contains oat milk.
b) Noor believes the milk pitcher contains almond milk.

Example 4: Default prompt

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor does not see her coworker swapping the milk.
Belief: Noor believes the milk pitcher contains oat milk.