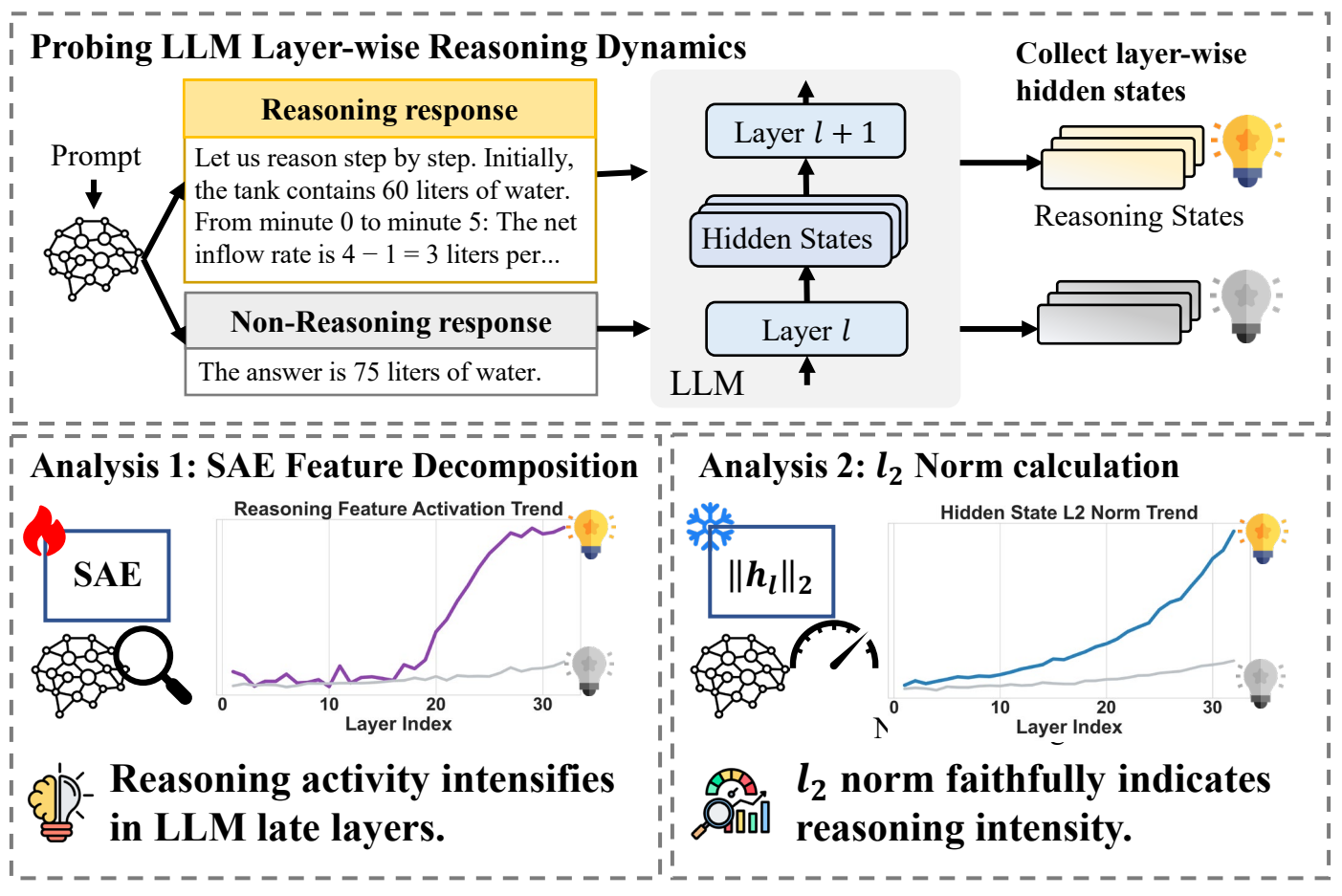
Probing LLM Layer-wise Reasoning Dynamics
Prompt
Reasoning response
Let us reason step by step. Initially, the tank contains 60 liters of water. From minute 0 to minute 5: The net inflow rate is 4 − 1 = 3 liters per...
Non-Reasoning response
The answer is 75 liters of water.
Layer $l+1$
Hidden States
Layer l
LLM
Collect layer-wise hidden states
Reasoning States
Analysis 1: SAE Feature Decomposition
SAE
Reasoning Feature Activation Trend
0
10
20
30
Layer Index
Reasoning activity intensifies in LLM late layers.
Analysis 2: l_2 Norm calculation
$\|h_l\|_2$
Hidden State L2 Norm Trend
0
10
20
30
Layer Index
l_2 norm faithfully indicates reasoning intensity.