

Fuel Gauge: Estimating Chain-of-Thought Length Ahead of Time in Large Multimodal Models



Making LMM Reasoning Predictable, Controllable, and Ready for New Applications
Yuedong Yang, Xiwen Wei, Mustafa Munir, Radu Marculescu



Paper

Why Does CoT Length Matter?

Reasoning LMMs generate *long and unpredictable* CoTs

- Memory fragmentation
- Over- and Under-thinking

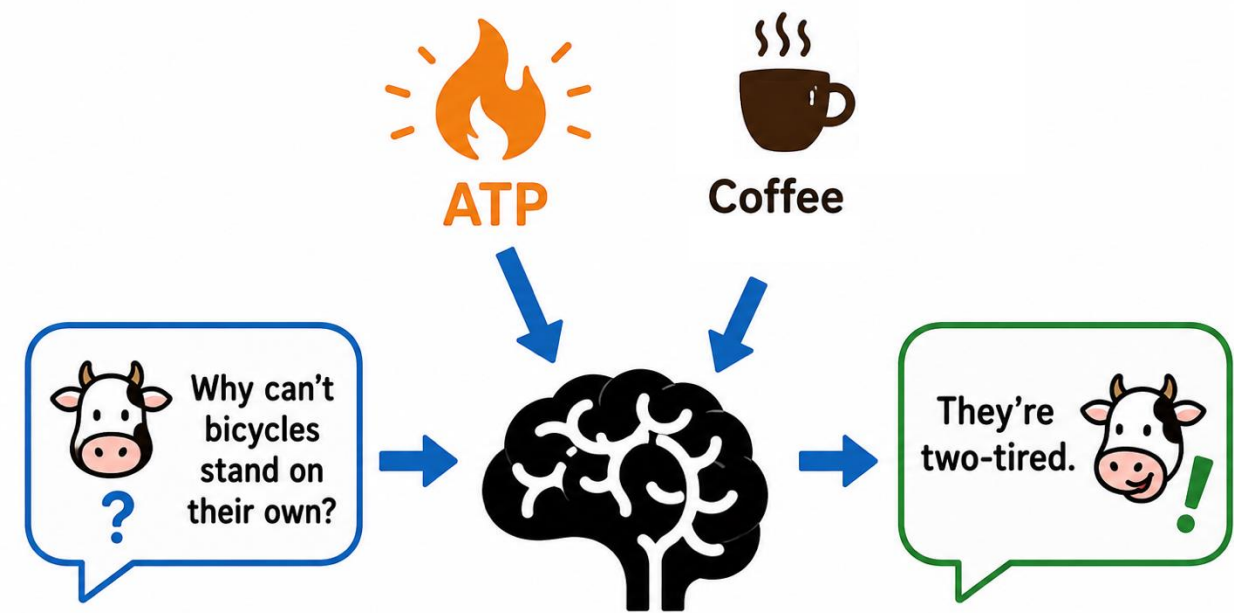
Question: "1 + 1 = ? (primary school)"

Chain of Thought

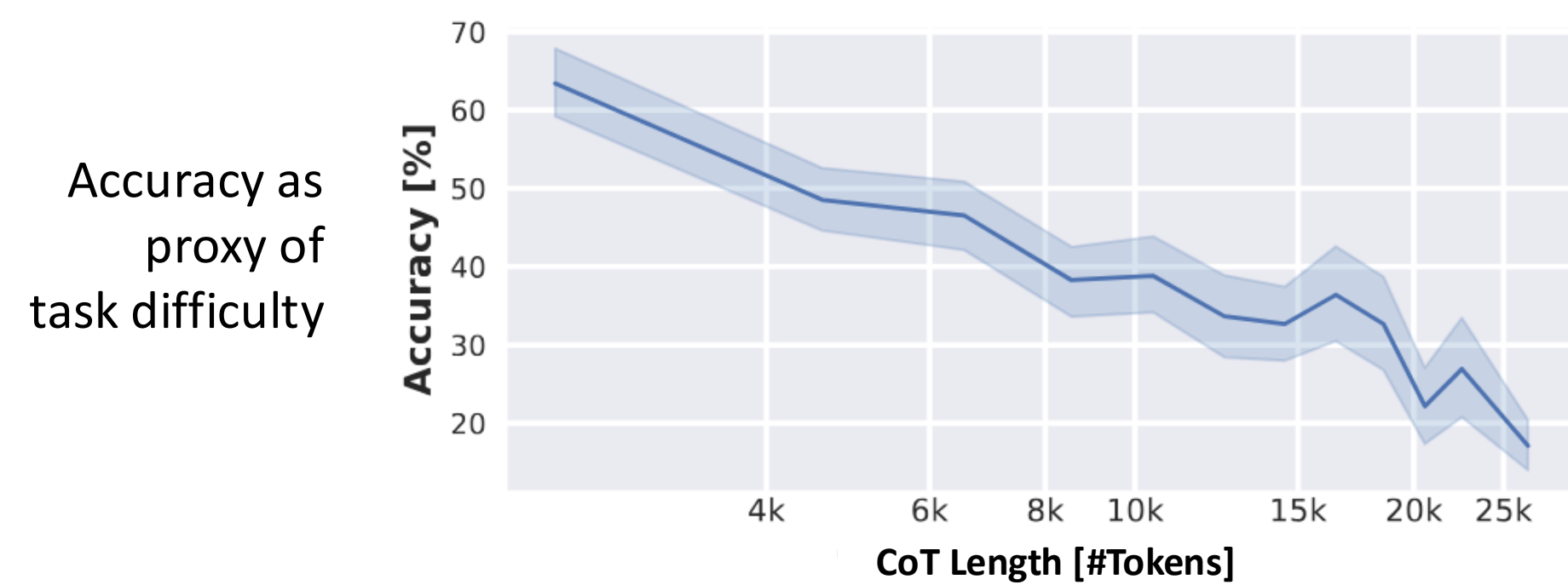
Okay, the user is asking 1 + 1 = ?. They mentioned it's for primary school. Let me think about primary school math problems. Usually, in first grade, basic addition and subtraction are taught. But wait, 1 + 1... That's straightforward. Maybe they're referring to a specific problem or an example? Wait, maybe there's a trick here. Sometimes in puzzles, numbers can be combined differently. For instance, if you consider the digits of the answer, but that seems unlikely. Alternatively, maybe it's a play on words where "1" and "1" are used in another way, like combining letters or something? But I don't see that in standard primary school math. Let me check if there's any variation. Some sources might present this as a common problem, but the answer is usually just 2. However, the user might have a different expectation. Since the question is from primary school, maybe the expected answer is 2. But I should confirm if there's any other possible interpretation. Alternatively, could it be a case where the numbers are in a sequence or a pattern? Like, adding two 1s to get another 1, which doesn't make sense. No, that's not right. So the most logical answer is 2. I'll go with that.

Answer: "1 + 1 = 2"

Human Thinking is "Predictable and Controllable"

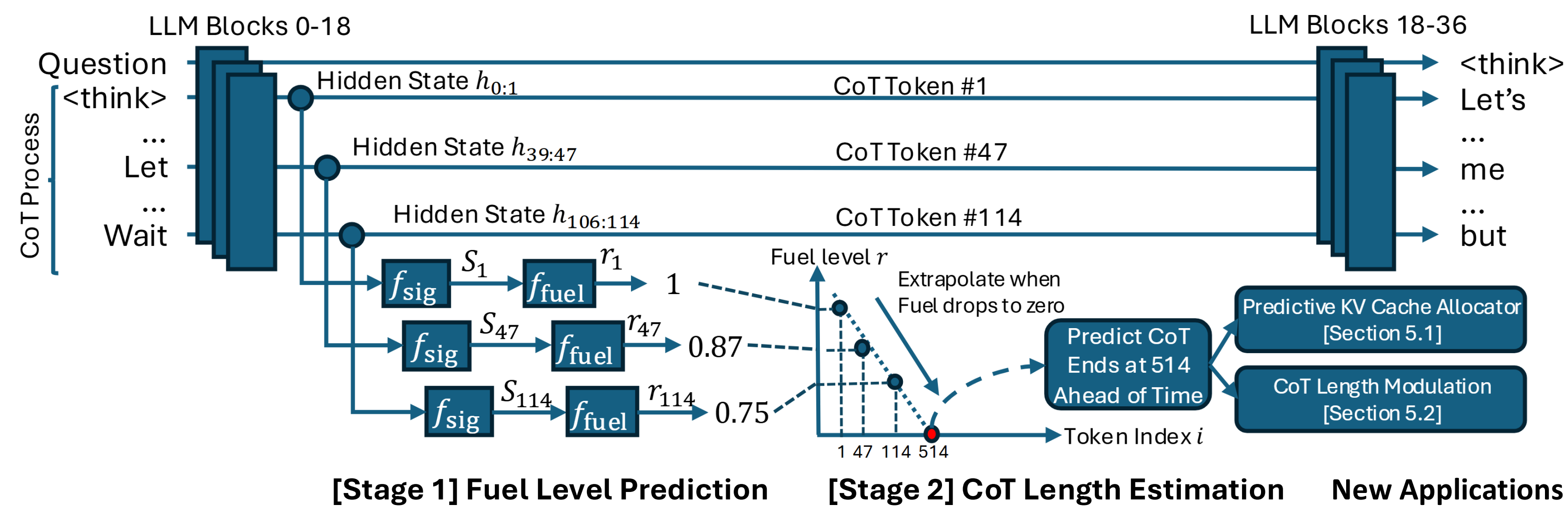


How about LMM Thinking?



Also "Predictable and Controllable"

Proposed Method: Fuel Gauge



[Stage 1] Fuel Level Prediction
Fuel level indicates the remaining "energy" supporting CoT, which drops from 1 to 0 as CoT progresses.

[Stage 2] CoT Length Estimation
Fit fuel level points with a line, predicting when it drops to zero, i.e., the end of CoT.

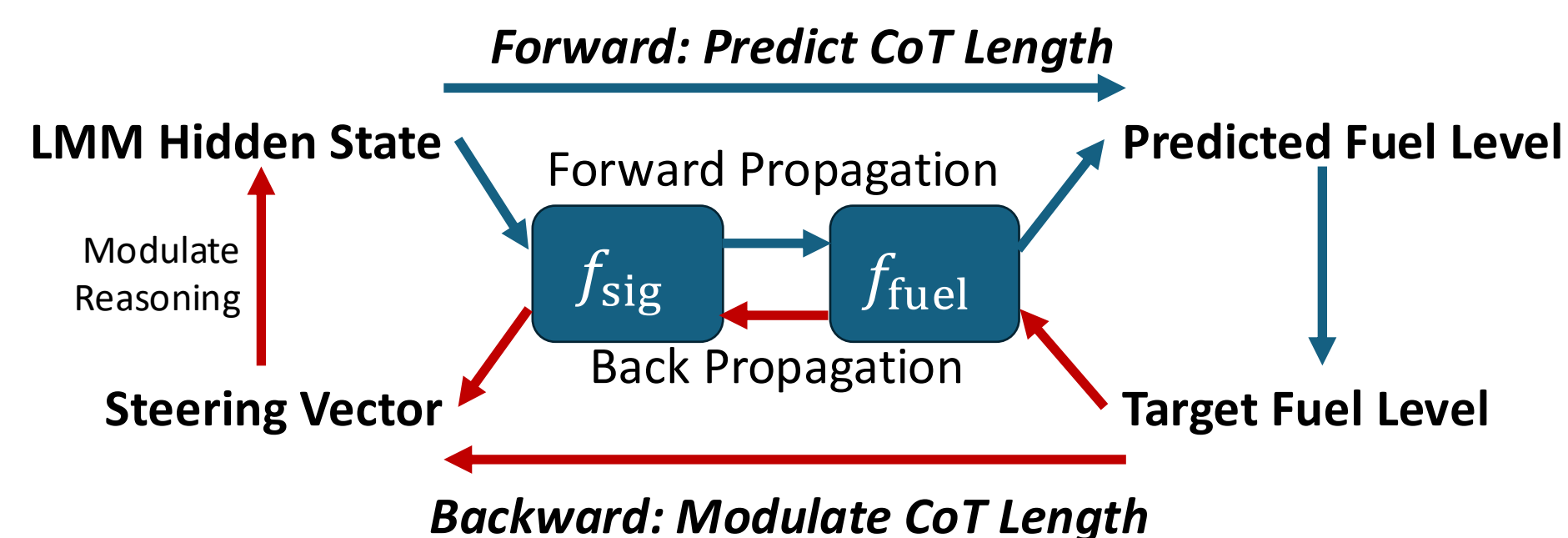
New Applications

Predictive KV Cache Allocator

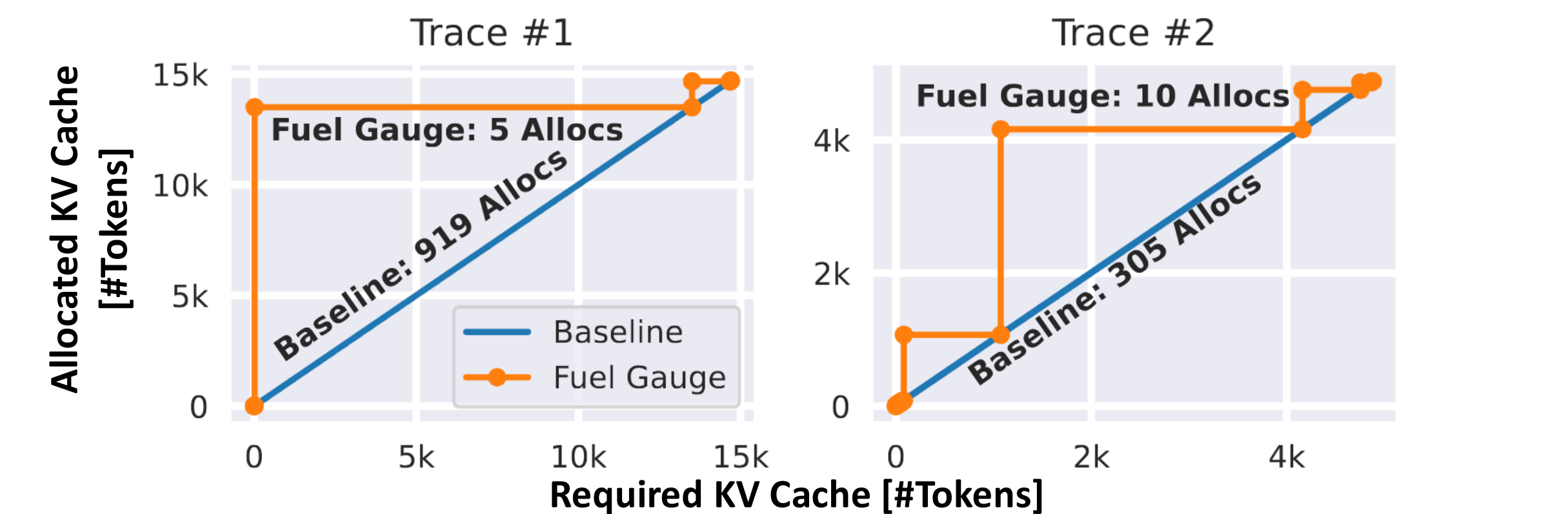
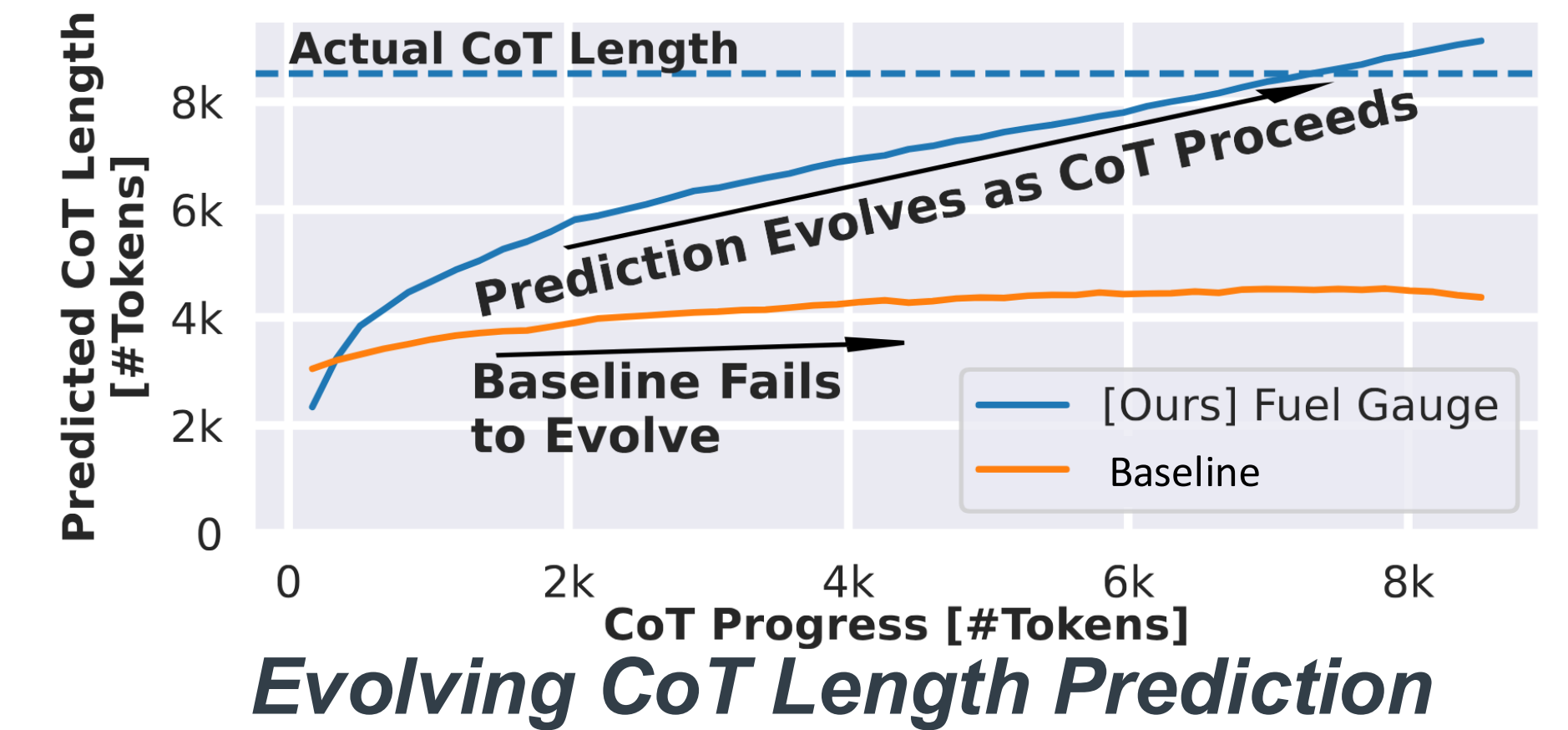
We can allocate memory based on the predicted CoT length. Update prediction and reallocate if allocated space is insufficient.

CoT Length Modulation

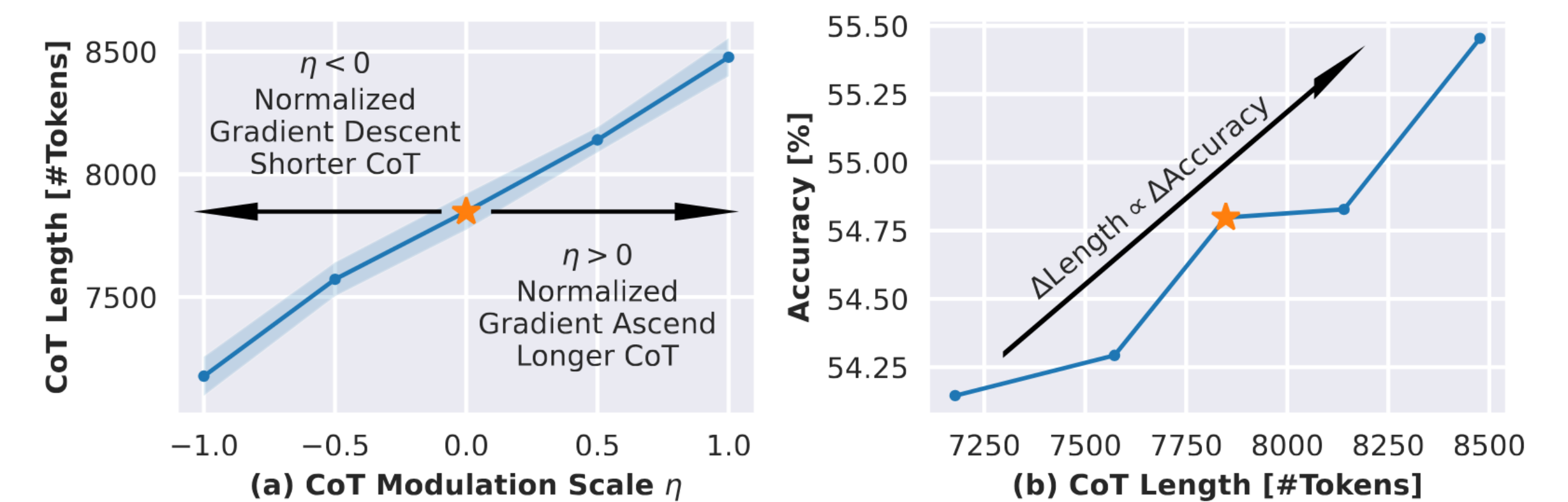
Since Fuel Gauge is fully differentiable, we can even control the reasoning behavior via gradient ascend or descend.



Evaluations



Significantly Reduced Memory Allocation Frequency



Quasi-linear Control over CoT Length and Quality