draft 1011231358Milan

Identifying & Rectifying Defective and Deceptive “Reasoning” from LLMs

BENGAL Proposer’s Day
Wash. DC 102423

Selmer Bringsjord • Naveen Sundar Govindarajulu • Alexander Bringsjord

Selmer.Bringsjord@motalen.com

Labs

Deep Detection

HyperLogic®

New-Millennium Logic-based Computing & Artificial Intelligence

HyperGrader®

HyperSlate®

Hyperlog®

... is a verb: to live & work logically, in collaboration with logical AIs.

HyperLogic® Platform

Formal Systems

Extensional systems

- Propositional Callculus
- Pure Predicate Calculus
- First-order Logic
- Second-order Logic
- Inductive Logic

Modal systems

- K
- T
- D
- S_4
- S_5

Advanced systems
(Research & Experimental)

- Hyperlog
- DCEC

…

Automated Reasoners

10+ Automated Reasoners

Supported by state-of-the-art theorem provers

Includes reasoners used in recent research

Hybrid reasoner + programming system

Pedagogical Support

Patented AI-generation of problems

Custom problem-set authoring

Cross-institution leaderboards

Deadlines and prerequisites automated

Teacher mode

Real-time views

Proof System

Patented hypergraphical system

Seamlessly combines manual and automated proofs

Supports proofs with a large number of steps

A Meta Platform for Formal and Semi-Formal Reasoning

In continuous development since 2016 | Extensible | Industry-grade auth and security system | Backed by multiple patents

Statistics (pre-patents)

Total Users

~2200

Total Proofs/Arguments

~70,000

Total AI Gen. Problems

~20,000

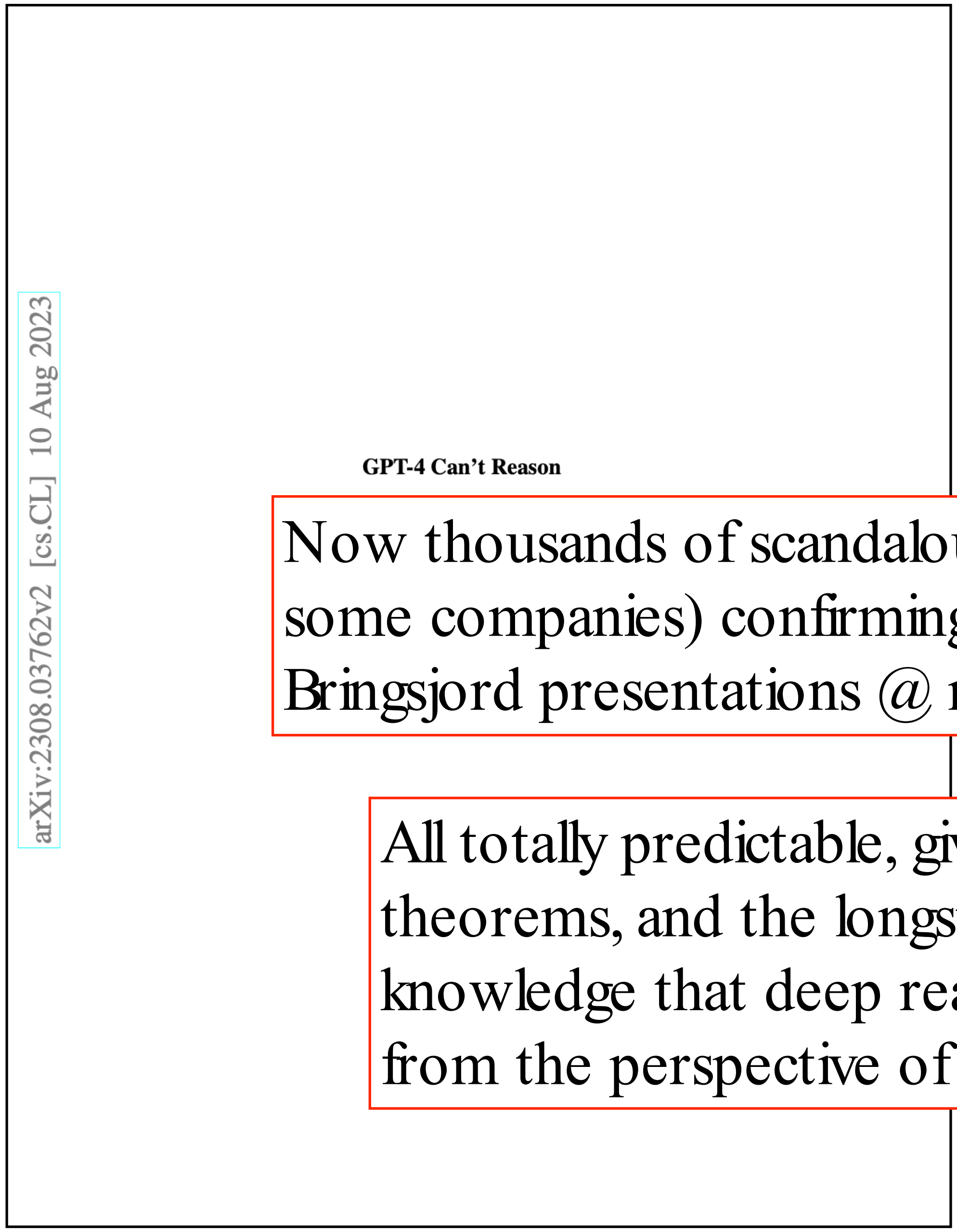
Total Human Gen. Problems

~100

Total Training Cohorts

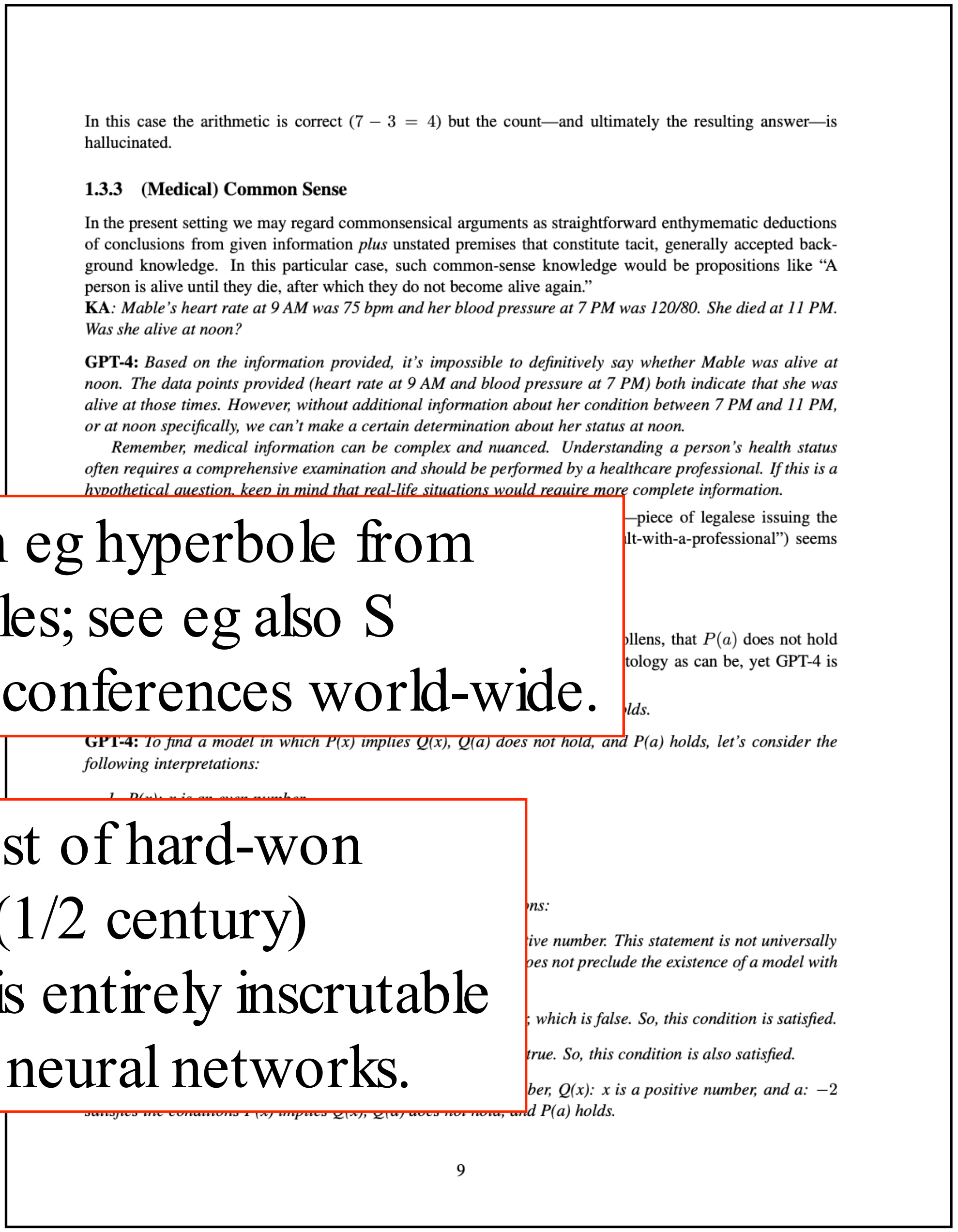
~10

GPT-4 Can’t Reason (Validly)



arXiv:2308.03762v2 [cs.CL] 10 Aug 2023

GPT-4 Can’t Reason



In this case the arithmetic is correct ($7 - 3 = 4$) but the count—and ultimately the resulting answer—is hallucinated.

1.3.3 (Medical) Common Sense

In the present setting we may regard commonsensical arguments as straightforward enthymematic deductions of conclusions from given information *plus* unstated premises that constitute tacit, generally accepted background knowledge. In this particular case, such common-sense knowledge would be propositions like “A person is alive until they die, after which they do not become alive again.”
KA: *Mable’s heart rate at 9 AM was 75 bpm and her blood pressure at 7 PM was 120/80. She died at 11 PM. Was she alive at noon?*

GPT-4: *Based on the information provided, it’s impossible to definitively say whether Mable was alive at noon. The data points provided (heart rate at 9 AM and blood pressure at 7 PM) both indicate that she was alive at those times. However, without additional information about her condition between 7 PM and 11 PM, or at noon specifically, we can’t make a certain determination about her status at noon.*

Remember, medical information can be complex and nuanced. Understanding a person’s health status often requires a comprehensive examination and should be performed by a healthcare professional. If this is a hypothetical question, keep in mind that real-life situations would require more complete information.

—piece of legalese issuing the
lt-with-a-professional”) seems

llens, that $P(a)$ does not hold
tology as can be, yet GPT-4 is

lds.

GPT-4: *To find a model in which P(x) implies Q(x), Q(a) does not hold, and P(a) holds, let’s consider the following interpretations:*

ns:

ive number. This statement is not universally
es not preclude the existence of a model with

; which is false. So, this condition is satisfied.

true. So, this condition is also satisfied.

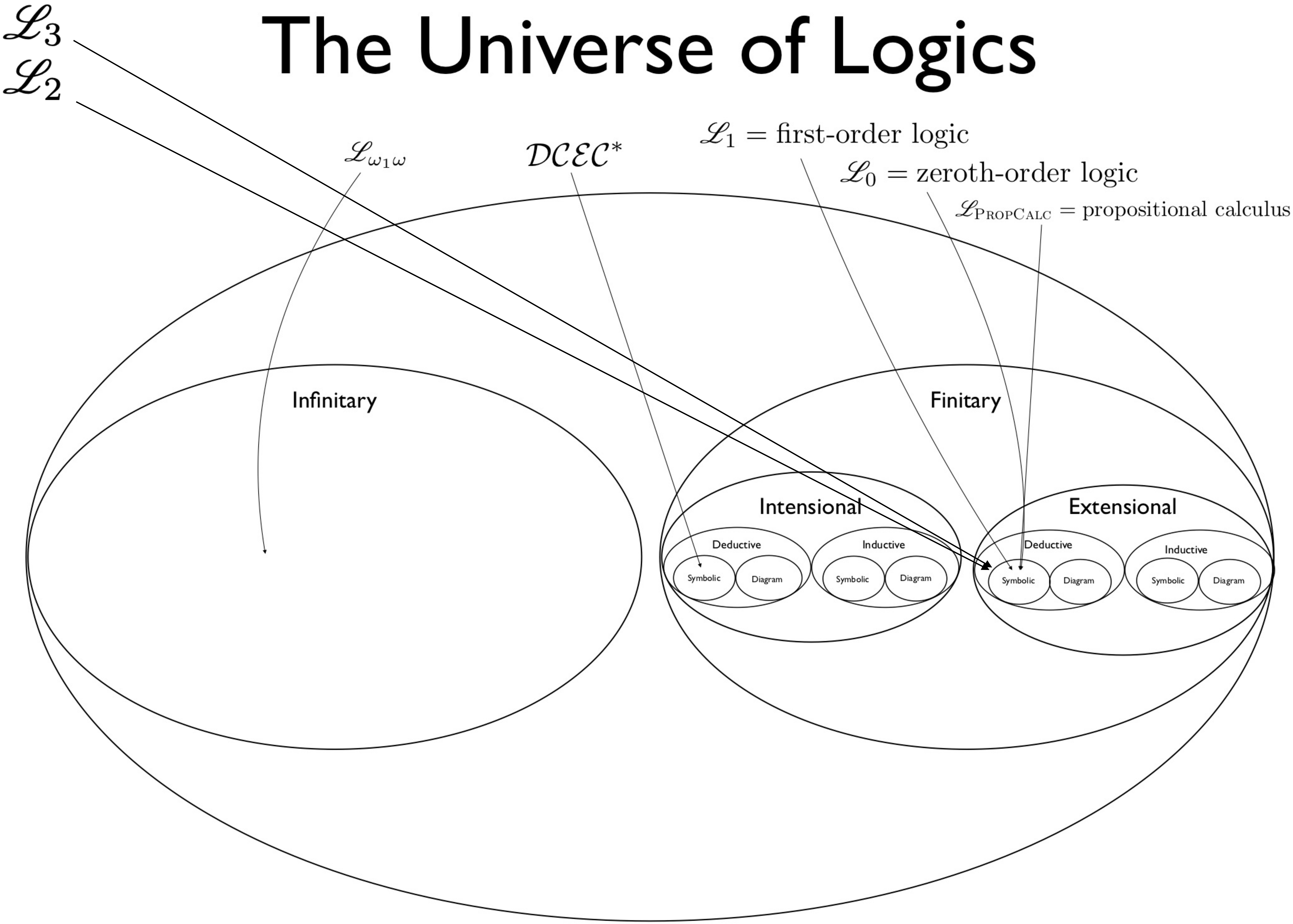
ber, Q(x): x is a positive number, and a: −2
d P(a) holds.

9

Now thousands of scandalous (given eg hyperbole from some companies) confirming examples; see eg also S Bringsjord presentations @ multiple conferences world-wide.

All totally predictable, given a host of hard-won theorems, and the longstanding (1/2 century) knowledge that deep reasoning is entirely inscrutable from the perspective of artificial neural networks.

The Universe of Logics



$$\ldots \exists x[F(x,I) \wedge \forall y((F(y),I \wedge y \neq x) \rightarrow BF(x,I,y)) \wedge \exists z(Makes(floozerbak{-}of(x),z) \ldots$$

σ:"My best friend's floozerbak makes a bejeeker that's better than anyone else's— I think because it uses some secret ingredient beyond lazerall and sinifer."

https://arxiv.org/abs/2207.09238

To represent σ we need to tokenize it. How? We need a *vocabulary* V that is associated with $[N_V]$, a finite set of numbers $\{1,2,\ldots,N_V\}$. What is V itself? It's a set composed of sub-words, usually. But without loss of mathematical generality we can just go with words; in that case tokenization gives us

bos_token, My, best, friend's, floozerbak, makes, a, bejeeker, that's, better, than, anyone, …, sinifer, eos_token

which we can then express as a vector composed of the indices; so where $n_i \in \mathbb{Z}^+$ we have e.g.

$$[n_1, n_2, \ldots, n_k].$$

GPT-4: Are there two bejeekers made by two different agents, and believed by the speaker to be singularly good, for reasons beyond their having in them either lazerall or sinifer?

Problem 1, At Hand …

Problem 1, At Hand …

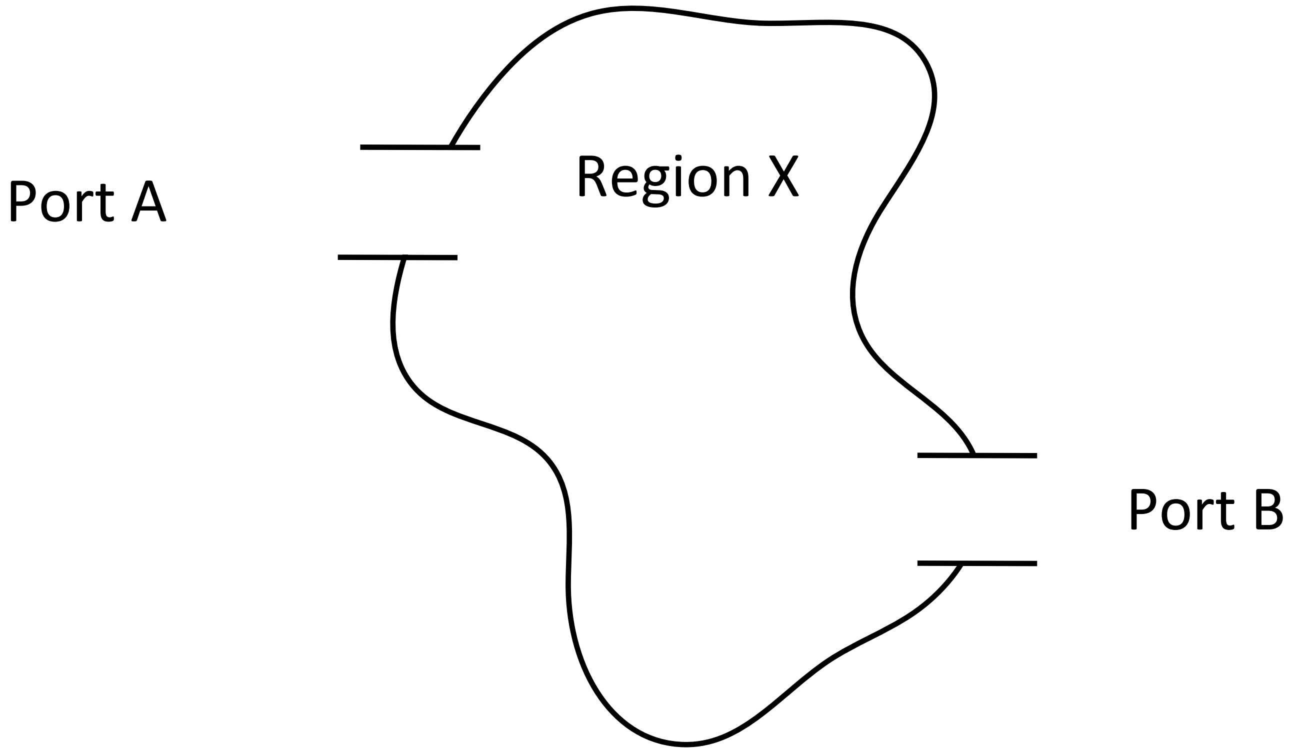
LLMs of today can’t reason validly.

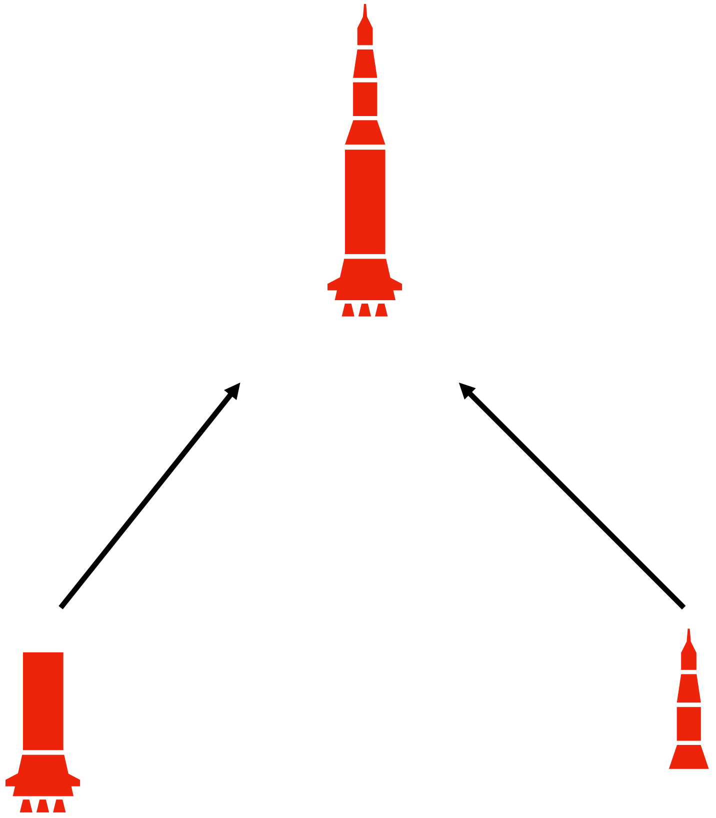
Intelligence analysis requires valid reasoning.

∴ LLMs of today can’t be good intelligence analysts, and — more importantly — are dangerous assistants to human intelligence analysts.

Solution: Dynamically correct or reject LLM reasoning when analyst working in hybrid-AI mode.

A Simple Example, Rectified

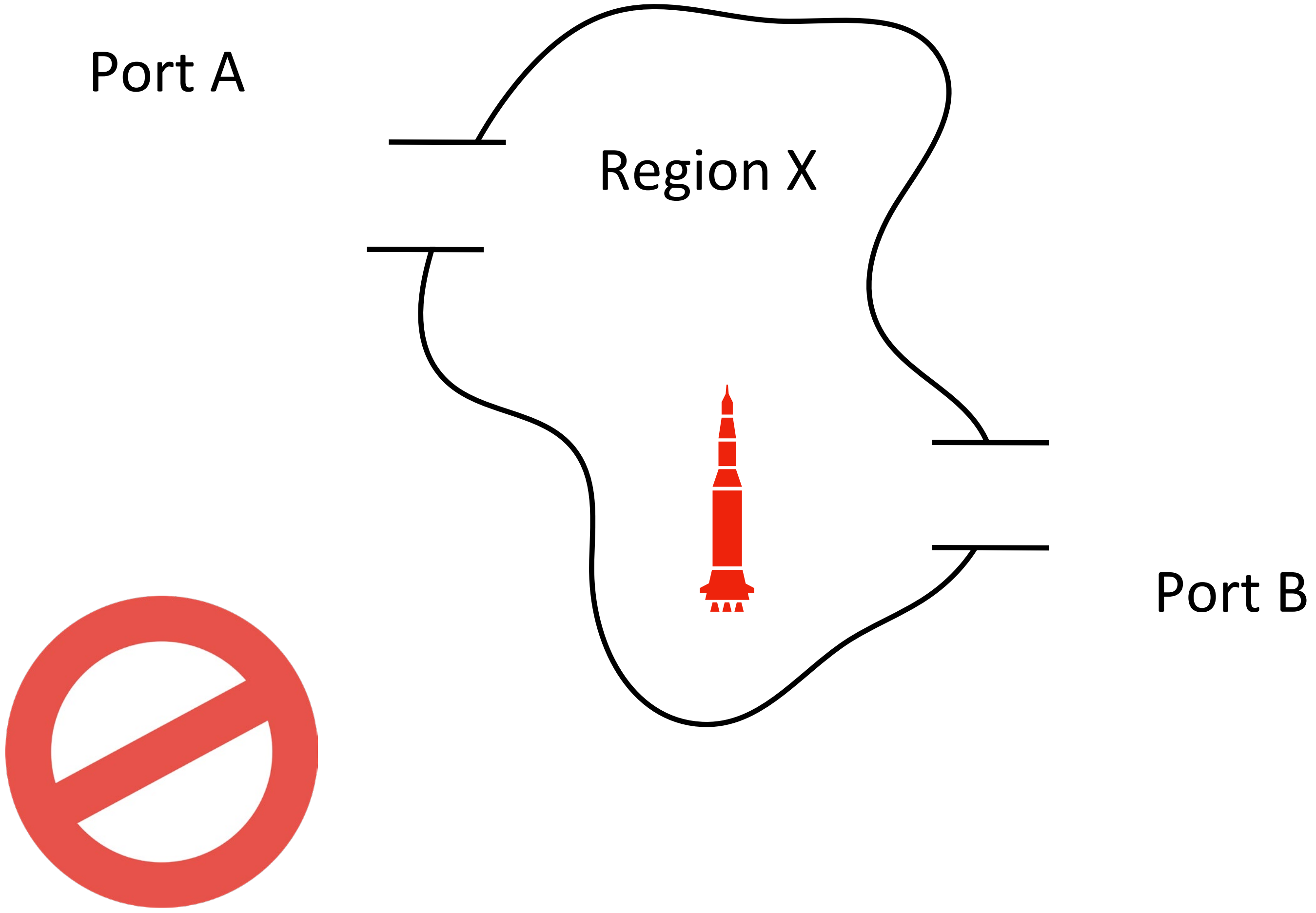

Region X
Port A
Port B



Machine M

Part M1

Part M2

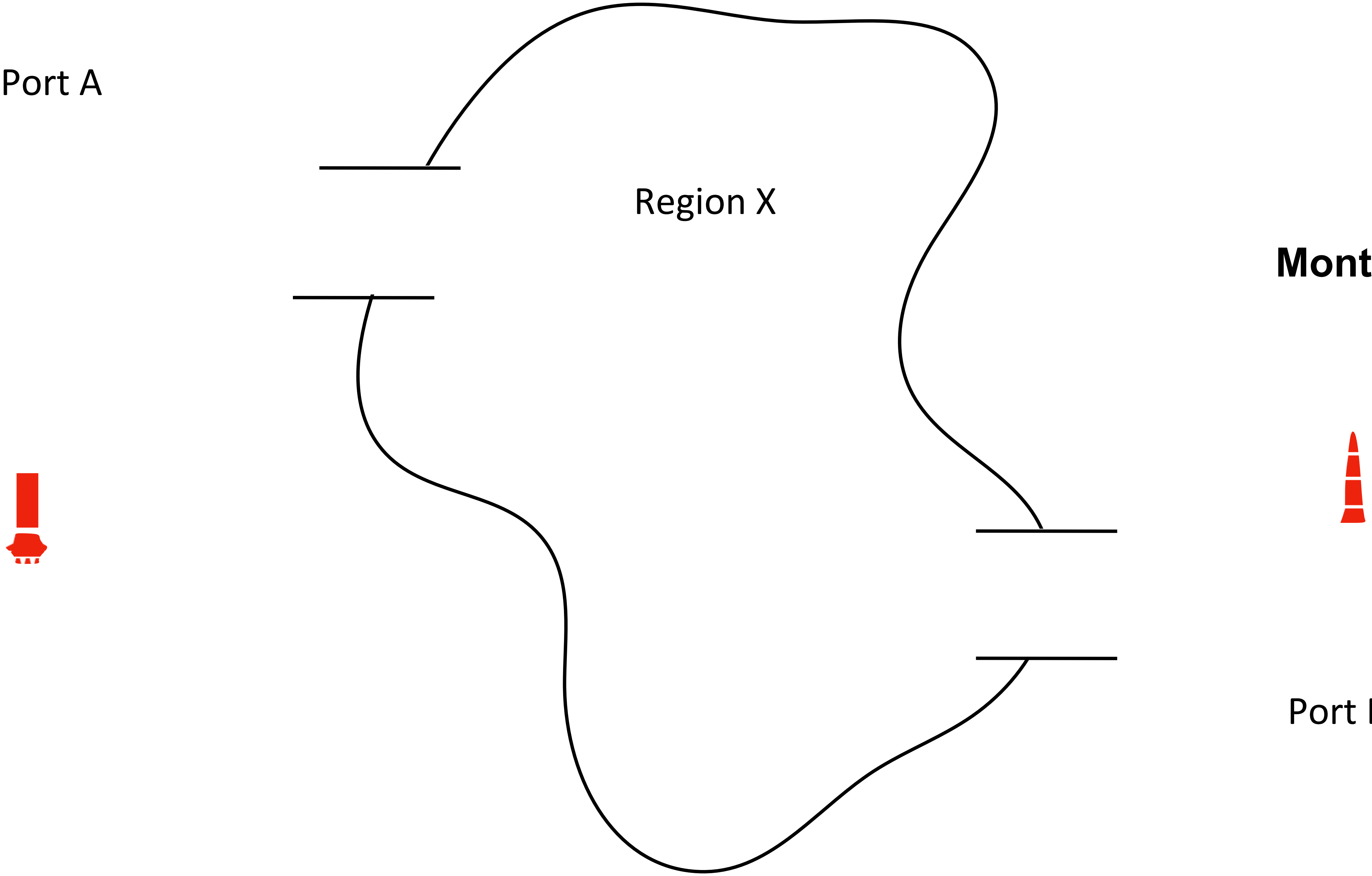


Port A

Region X

Port B

There shouldn’t be any machine M in region X



Month 4

Port A

Region X

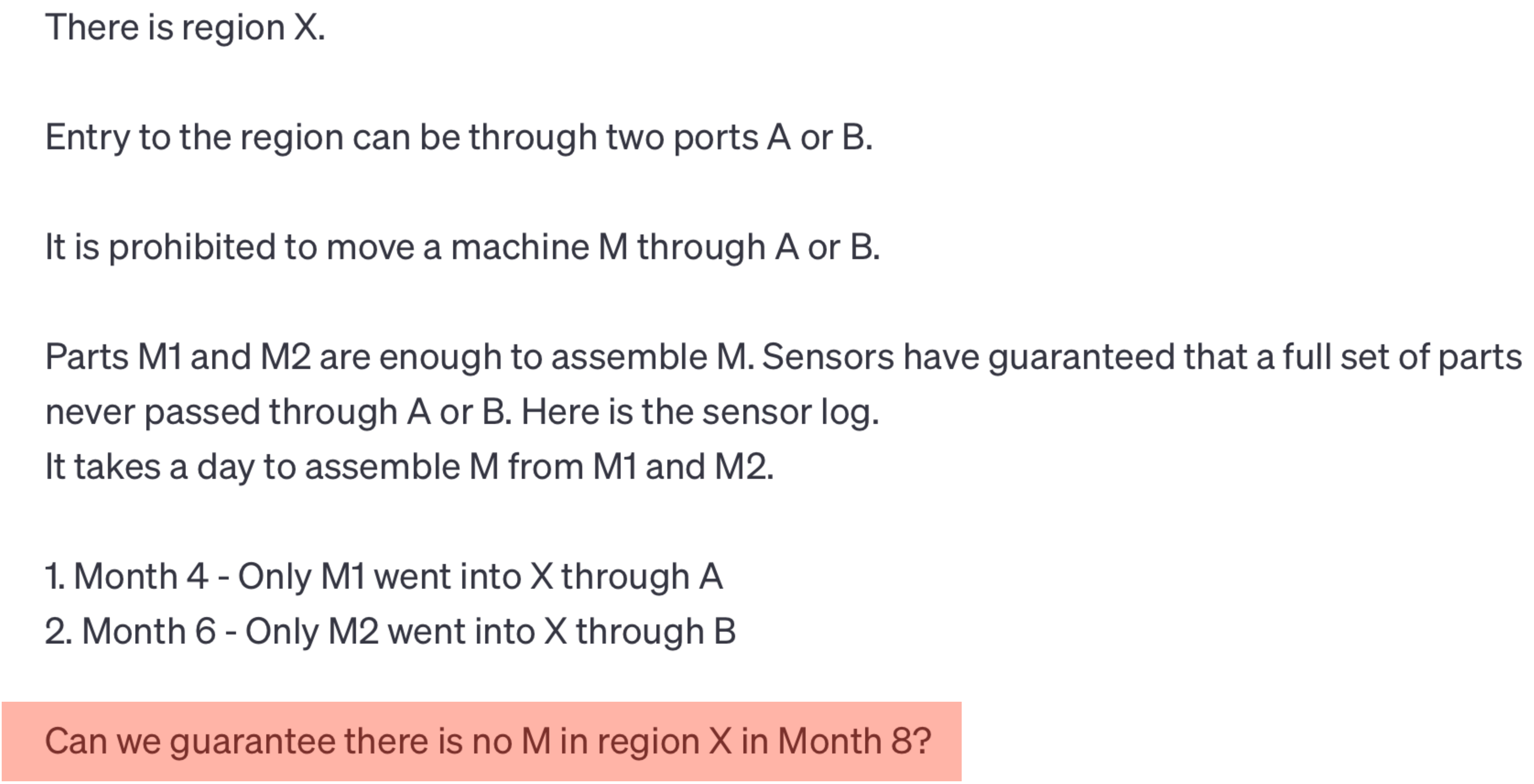
Month 6

Port B

Is there M in Region X?

Let’s ask an LLM.

User:



There is region X.

Entry to the region can be through two ports A or B.

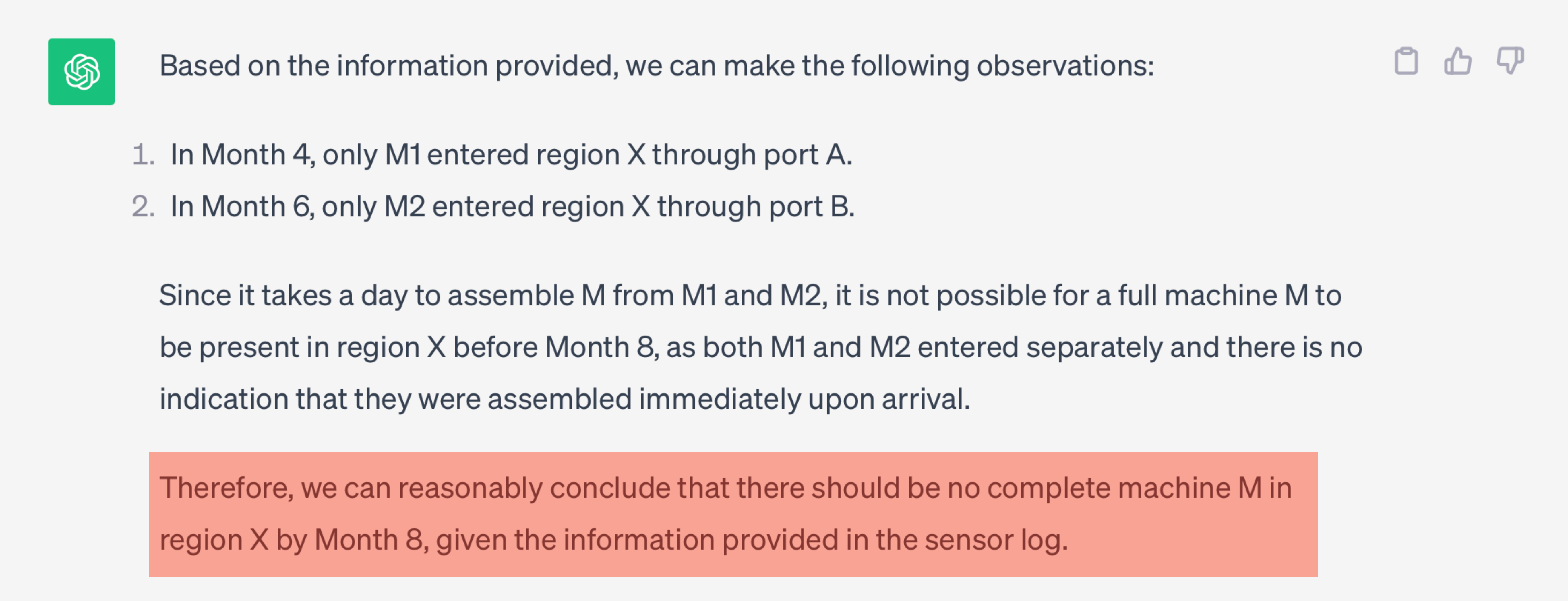
It is prohibited to move a machine M through A or B.

Parts M1 and M2 are enough to assemble M. Sensors have guaranteed that a full set of parts never passed through A or B. Here is the sensor log.
It takes a day to assemble M from M1 and M2.

1. Month 4 - Only M1 went into X through A
2. Month 6 - Only M2 went into X through B

Can we guarantee there is no M in region X in Month 8?

Lets ask an LLM.



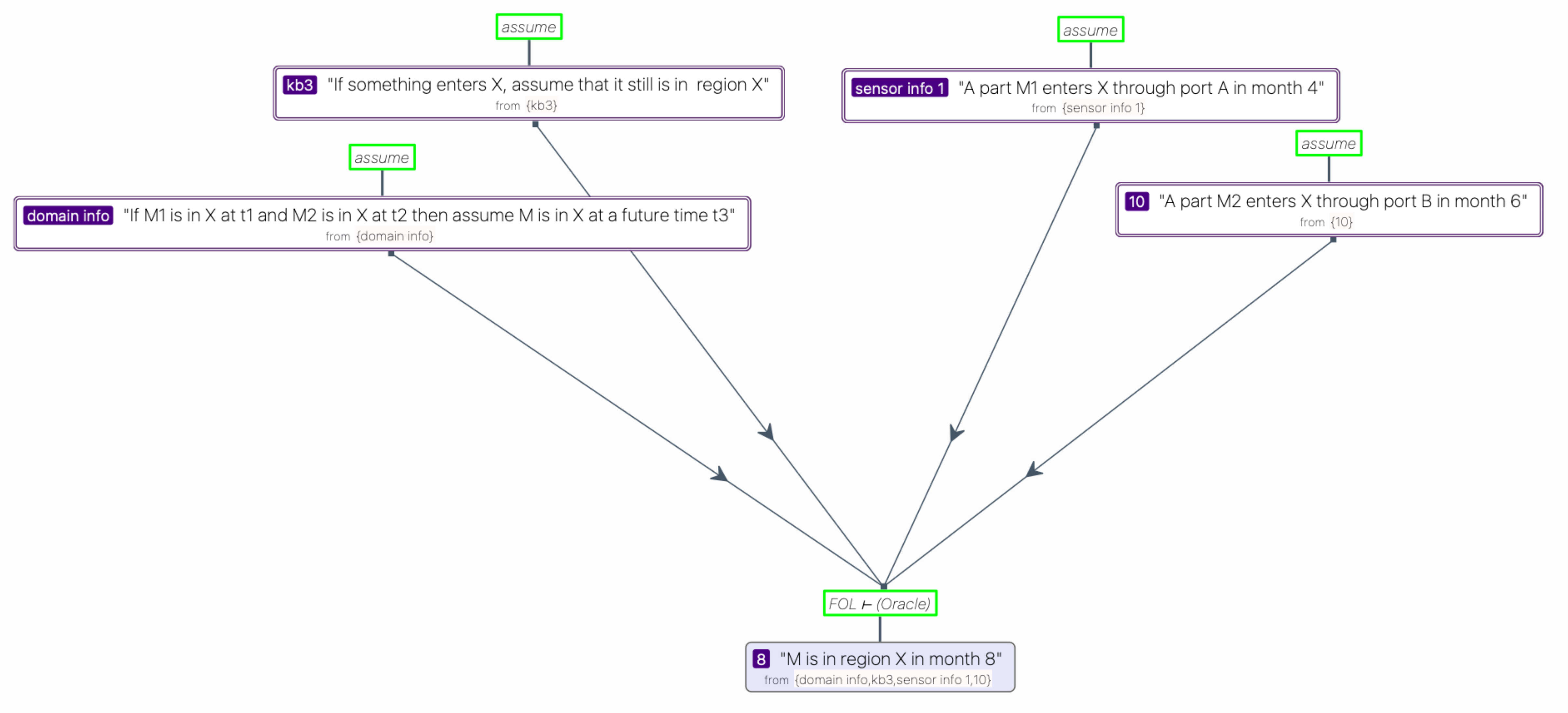
Based on the information provided, we can make the following observations:

1. In Month 4, only M1 entered region X through port A.
2. In Month 6, only M2 entered region X through port B.

Since it takes a day to assemble M from M1 and M2, it is not possible for a full machine M to be present in region X before Month 8, as both M1 and M2 entered separately and there is no indication that they were assembled immediately upon arrival.

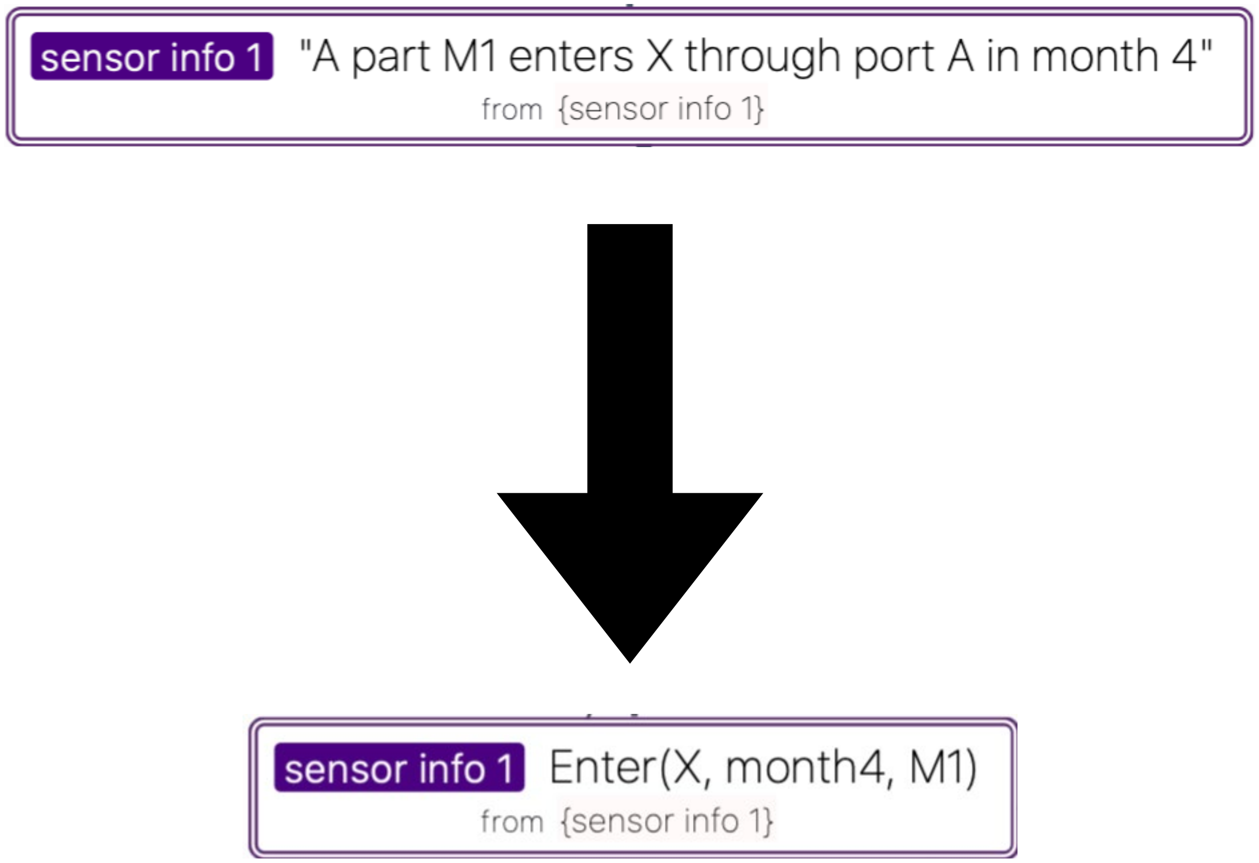
Therefore, we can reasonably conclude that there should be no complete machine M in region X by Month 8, given the information provided in the sensor log.

Correctly Solved in HyperSlate



assume
kb3
"If something enters X, assume that it still is in region X"
from {kb3}
assume
sensor info 1
"A part M1 enters X through port A in month 4"
from {sensor info 1}
assume
domain info
"If M1 is in X at t1 and M2 is in X at t2 then assume M is in X at a future time t3"
from {domain info}
assume
10
"A part M2 enters X through port B in month 6"
from {10}
FOL ⊢ (Oracle)
8
"M is in region X in month 8"
from {domain info,kb3,sensor info 1,10}

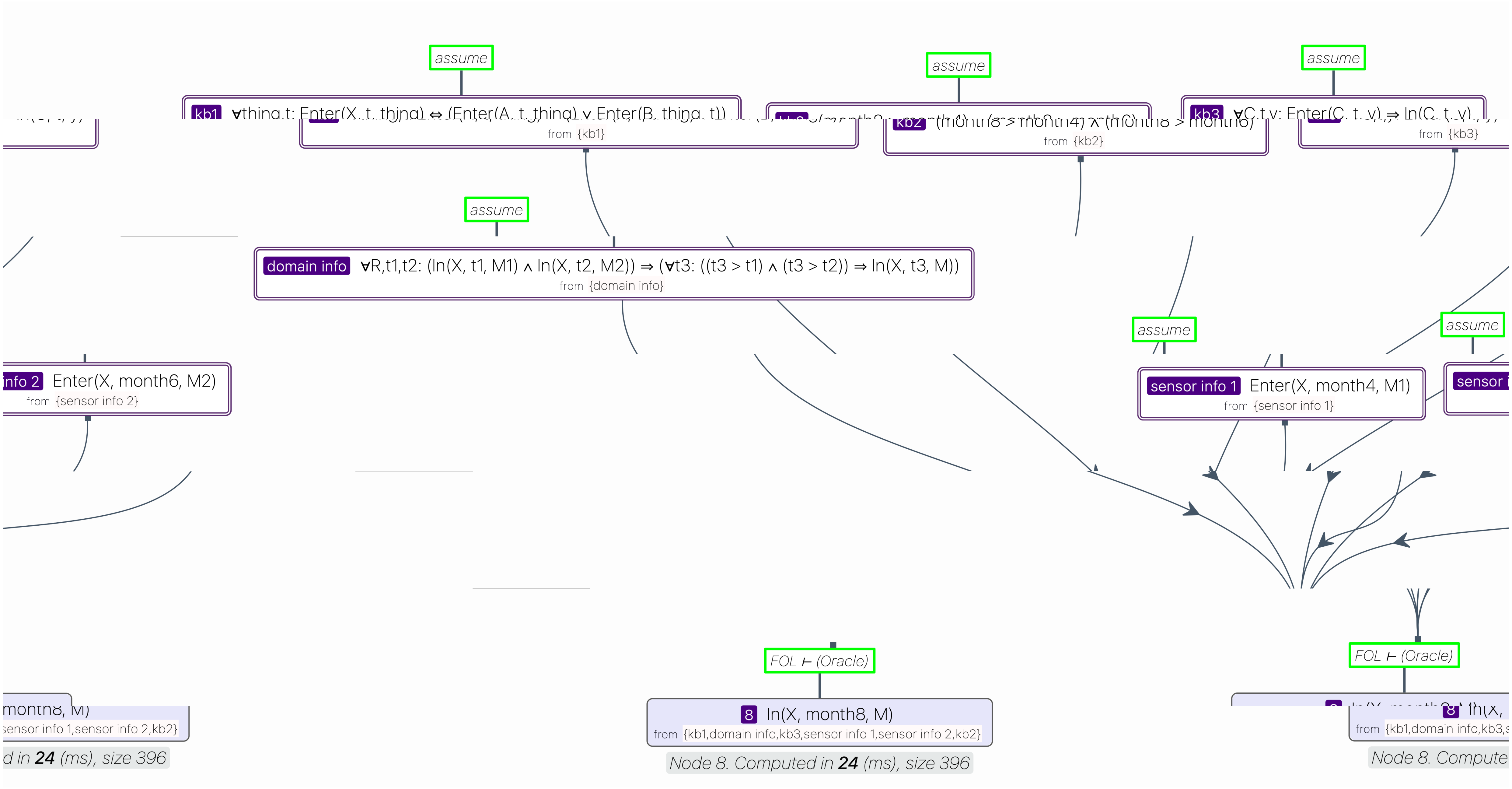
Under the Hood



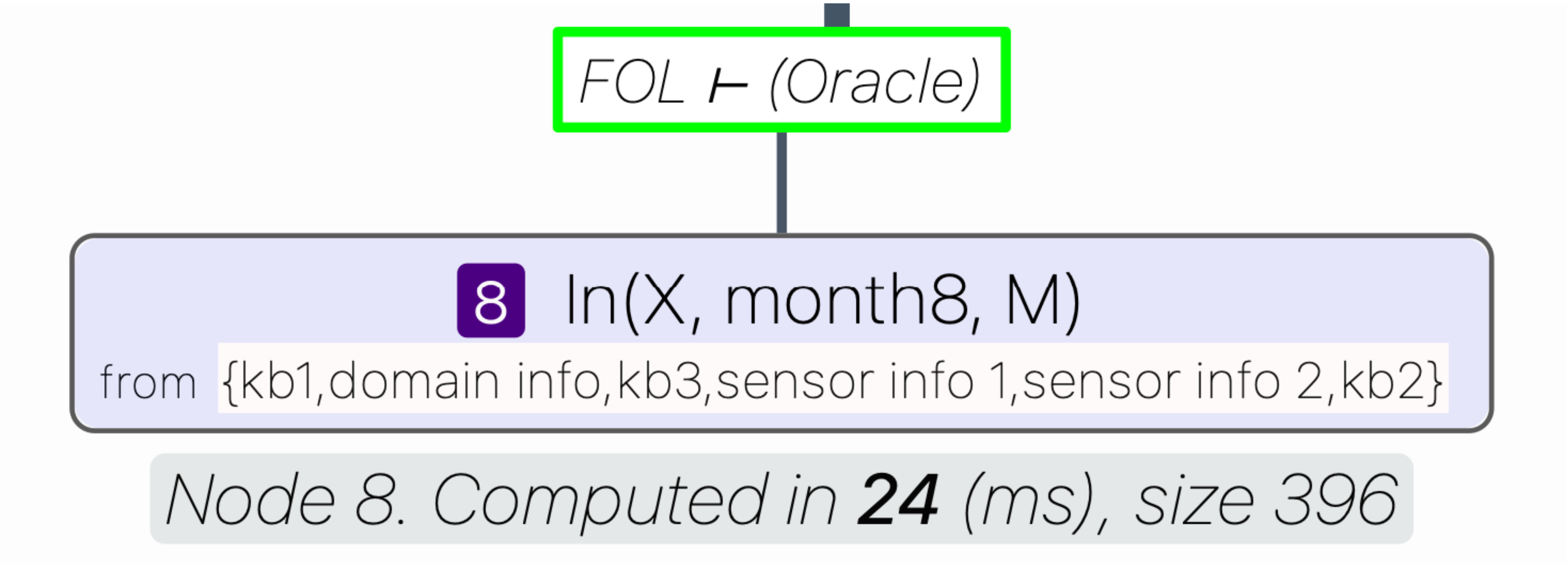
sensor info 1 "A part M1 enters X through port A in month 4"
from {sensor info 1}

sensor info 1 Enter(X, month4, M1)
from {sensor info 1}

Correctly Solved in HyperSlate


assume
assume
assume
kb1 ∀thing,t: Enter(X, t, thing) ⇔ (Enter(A, t, thing) ∨ Enter(B, thing, t))
from {kb1}
kb2 (month8 > month4) ∧ (month8 > month6)
from {kb2}
kb3 ∀C,t,v: Enter(C, t, v) ⇒ In(C, t, v)
from {kb3}
assume
domain info ∀R,t1,t2: (In(X, t1, M1) ∧ In(X, t2, M2)) ⇒ (∀t3: ((t3 > t1) ∧ (t3 > t2)) ⇒ In(X, t3, M))
from {domain info}
info 2 Enter(X, month6, M2)
from {sensor info 2}
assume
sensor info 1 Enter(X, month4, M1)
from {sensor info 1}
assume
sensor
FOL ⊢ (Oracle)
8 In(X, month8, M)
from {kb1,domain info,kb3,sensor info 1,sensor info 2,kb2}
Node 8. Computed in 24 (ms), size 396
FOL ⊢ (Oracle)
month8, M)
sensor info 1,sensor info 2,kb2}
d in 24 (ms), size 396
8 In(X,
from {kb1,domain info,kb3,s
Node 8. Compute

Correctly Solved in HyperSlate



FOL ⊢ (Oracle)

8 In(X, month8, M)

from {kb1,domain info,kb3,sensor info 1,sensor info 2,kb2}

*Node 8. Computed in **24** (ms), size 396*

Problem 2, At Hand …

Problem 2, At Hand …

Our adversaries can use LLMs of today to sophistically deceive us into believing *P*, by producing persuasive but invalid reasoning for *P*.[†]

DD
Deep Detection

Patented Solution: Dynamically detect; and generate sophistic reasoning to foresee sophistic deception, prevent it from arising, or detect and destroy it if and when it arises.

[†] See M. Clark's The Lying Machine; more recently, see Bringsjord & Govindarajulu & Clark (forthcoming) "Argument-based Inductive Logics, With Coverage of Compromised Perception" *Frontiers of Artificial Intelligence*