Natural Language Processing

CSE 447

Grand Challenges in LLM Reasoning

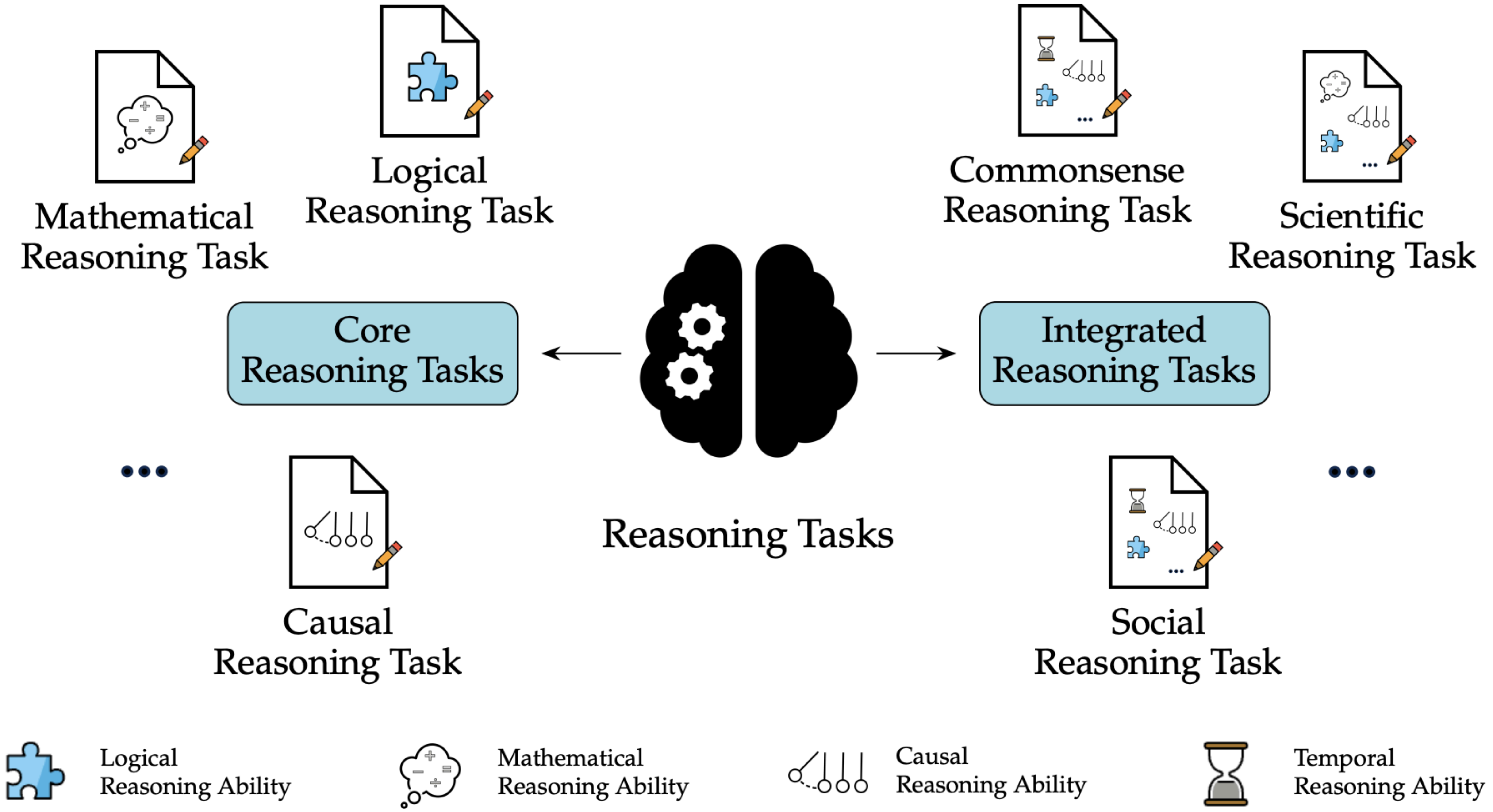
Lecturer: Melanie Sclar

Some slides from Hyunwoo Kim

Outline

- **What does LLM reasoning encompass?**
- **Grand Challenges:**
  - How do we **measure** LLMs' reasoning skills?
    - Striving to improve measurement practices: Theory of Mind as a case study.
    - Quantifying memorization vs generalization through rationale-based reasoning
  - How do we **improve** LLMs' reasoning skills?
    - Training techniques, or training with better data.
    - Chain of Thought is not a holy grail. Inference-time algorithms.

What does LLM reasoning encompass?



From Mondorf and Plank, 2024. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey.

How can reasoning be measured?

Evaluation Method	Advantages	Disadvantages
Conclusion-based evaluation	Allows for controlled setups Provides metrics for comparison Easy to automate and scale Easy to reproduce	Limited insights Less reliable
Rationale-based evaluation	Offers more nuanced insights More robust in certain scenarios	Difficult to automate and scale Might require expert interpretation
Interactive evaluation	Highly flexible Customizable to model behavior	Expensive Difficult to automate and scale Less standardized and reproducible
Mechanistic evaluation	Identifies features or circuits responsible for specific behaviors Supports direct interventions on model internals	Findings may not generalize across tasks or models Results may be hard to interpret Compute-intensive

From Mondorf and Plank, 2024. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey.

Faithful Reasoning Evaluation

Rationale-based *evaluation for measuring generalization vs memorization in mathematical reasoning*

Faith and Fate: Limits of Transformers on Compositionality

NeurIPS 2023 (Spotlight)



Nouha Dziri* Ximing Lu* Melanie Sclar*

Lorraine Li† Liwei Jiang† Bill Yuchen Lin† Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, Yejin Choi

Are LLMs truly reasoning, or are they memorizing from training data?

How can we characterize model performance with respect to properties of each task sample?

When we see models solving a seemingly difficult question, what can we correlate it to? How do we characterize model errors?

Measuring & characterizing compositionality

We need to decouple our analysis from pre-training data (inaccessible to us).

- Multi-step reasoning tasks we use: long-form multiplication, a dynamic programming task & Einstein's puzzle (word logic puzzle)

Math/word logic reasoning problems are ideal: infinite data to be generated that the model can't have possibly seen in its entirety!

- Our method: We train models (GPT3) to generate step-by-step solutions for each task, and view their solutions as **computation graphs.** We can then compare them to ground truth graphs!

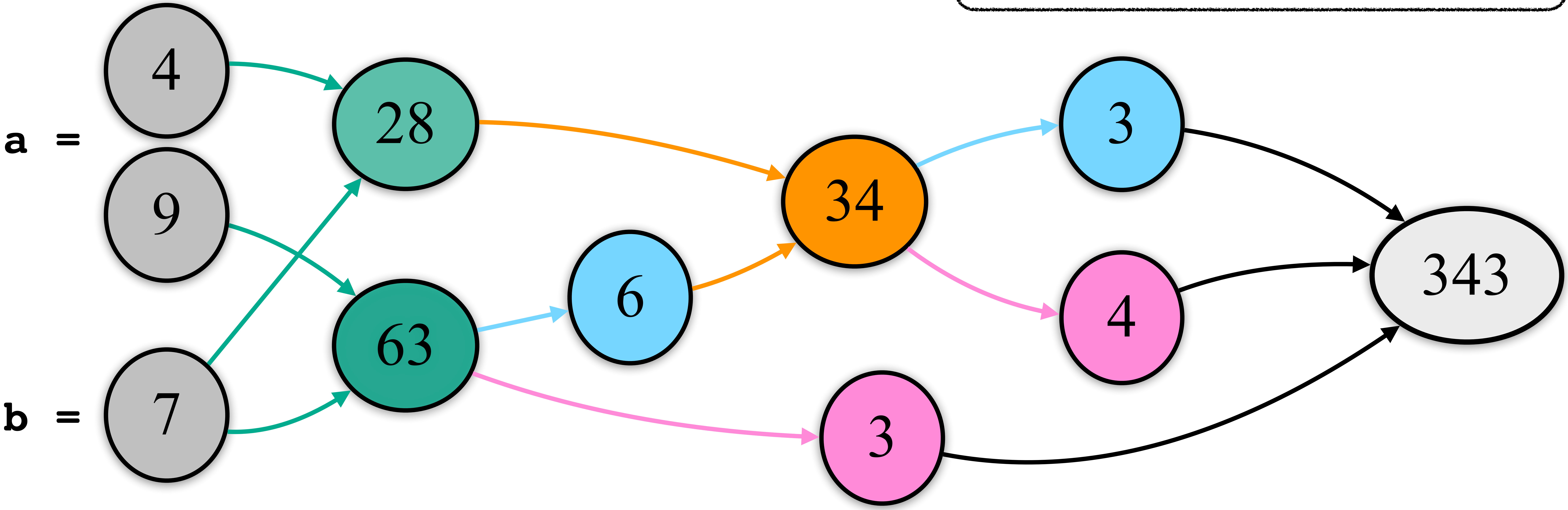
Computation graph for 49 x 7

multiply 1-digit | mod 10 | carry | sum | concat

```
function multiply (a[1:p], b[1:q]):
 for i = q to 1
    carry = 0
    for j = p to 1
    t = a[j] * b[i]
    t += carry  (only if j != p)
    digits[j] = t mod 10
    carry = t // 10
    summands[i] = digits
```

$$\textit{product} = \sum_{i=1}^{q} \texttt{summands[q+1-i]} \cdot 10^{i-1}$$

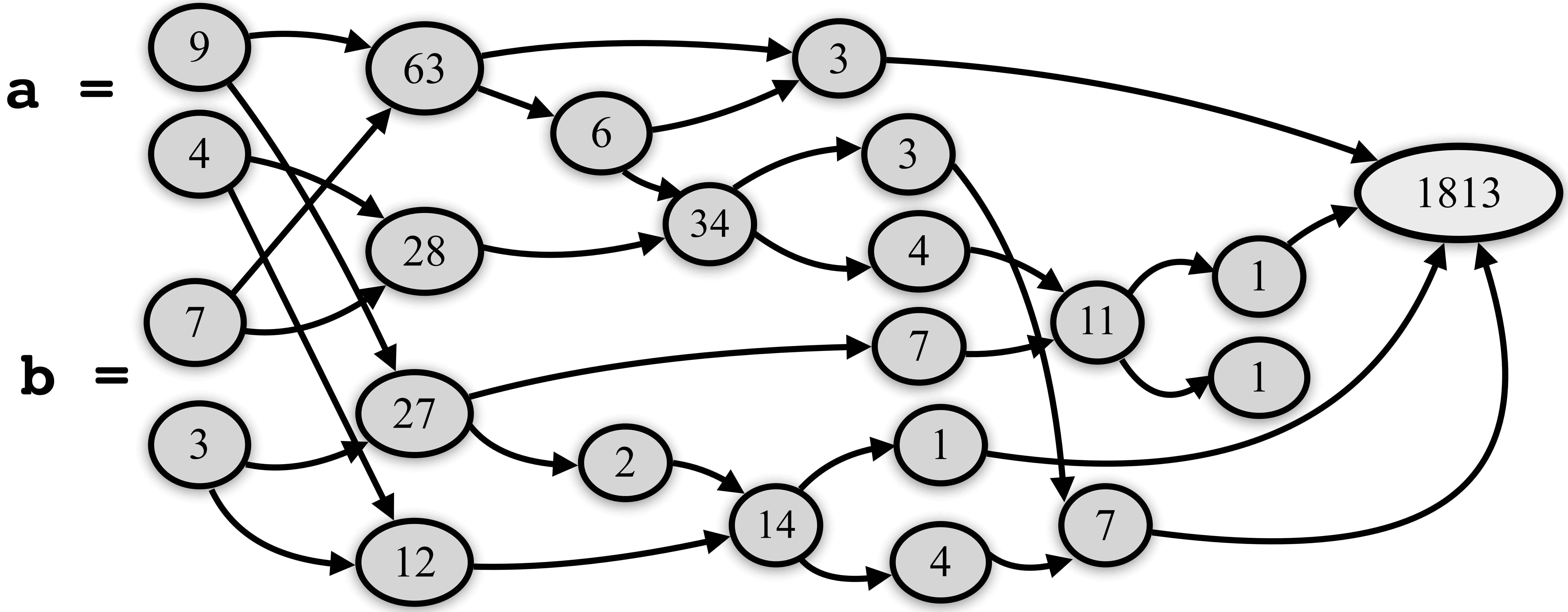
```
  return product
```



How can we characterize model performance with respect to properties of each task sample?

When we see models solving a seemingly difficult question, what can we correlate it to?
How do we characterize model errors?

Information Gain Explains Where Transformers Partially Excel

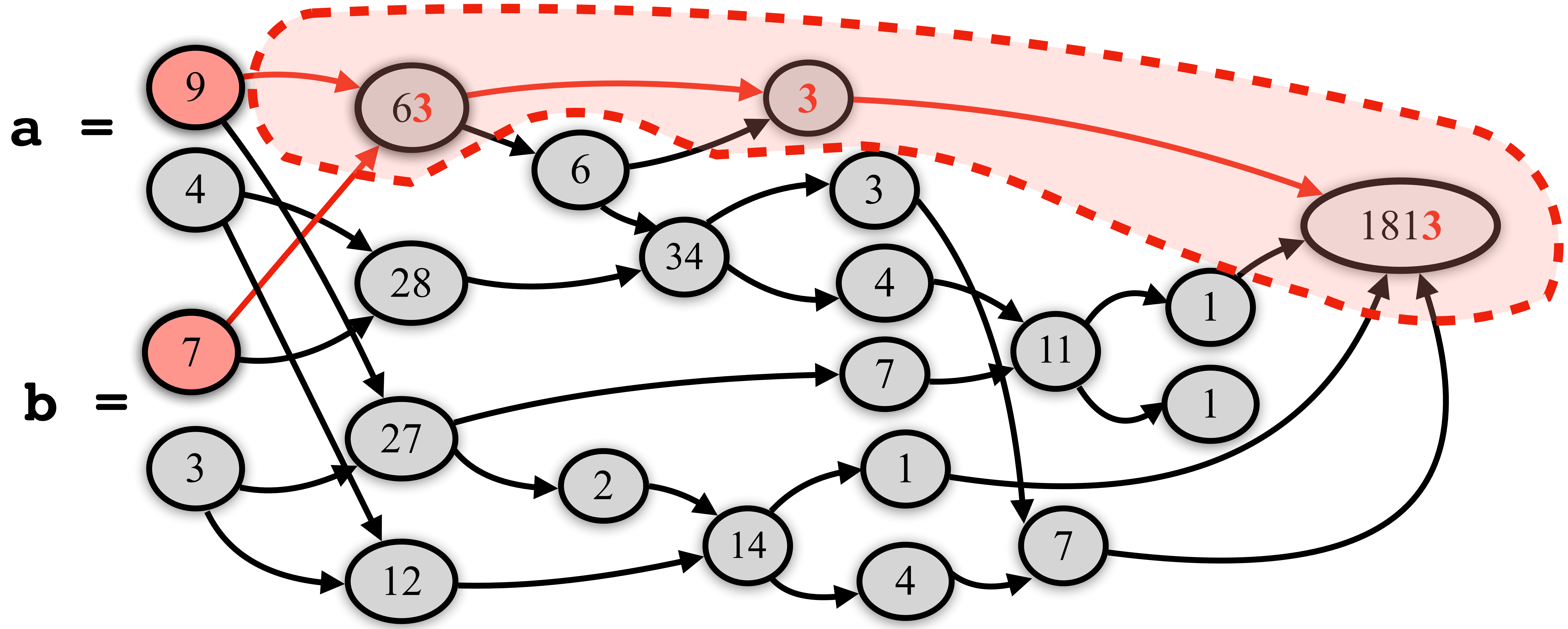


Information Gain Explains Where Transformers Partially Excel

49
37

343
147

= 1813



Information Gain Explains Where Transformers Partially Excel

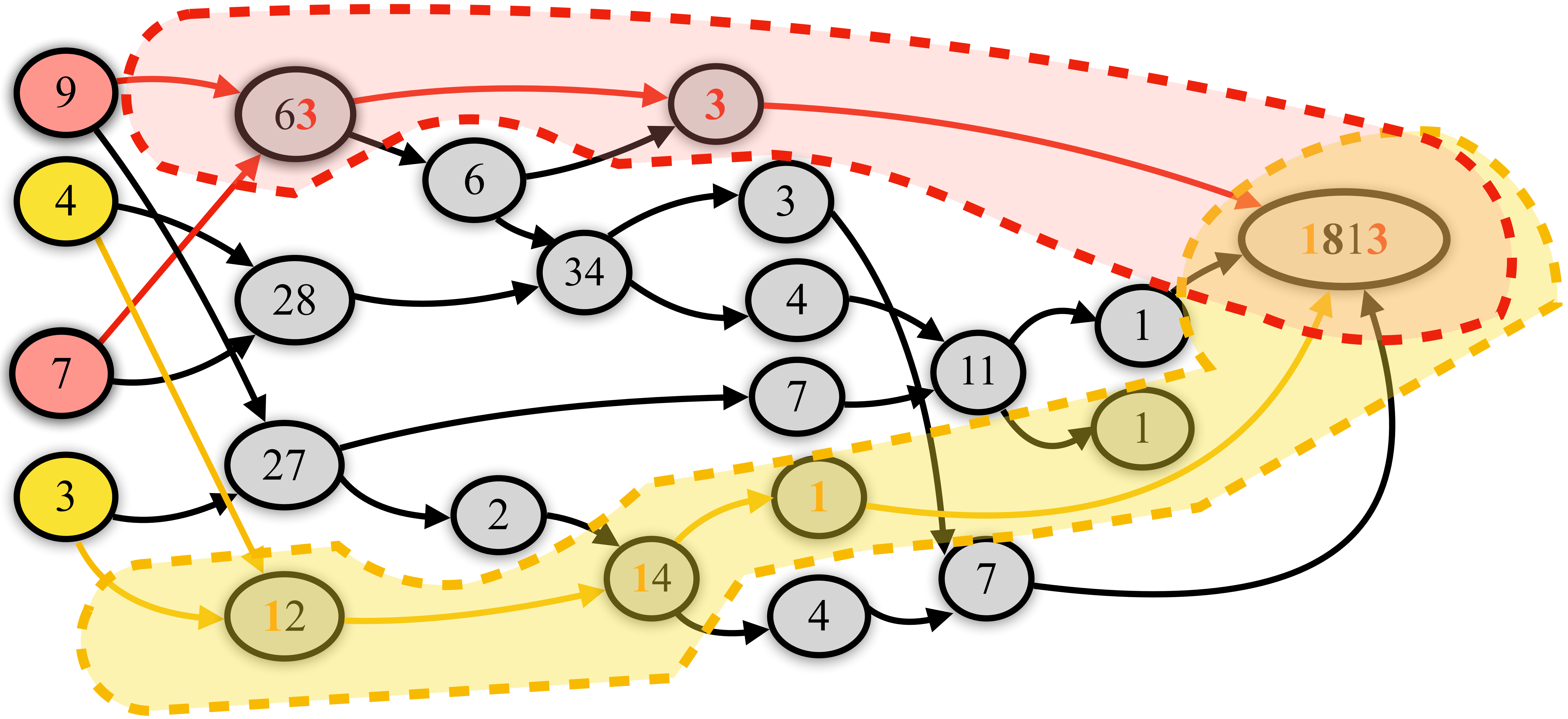
49
37

343
147

= 1813

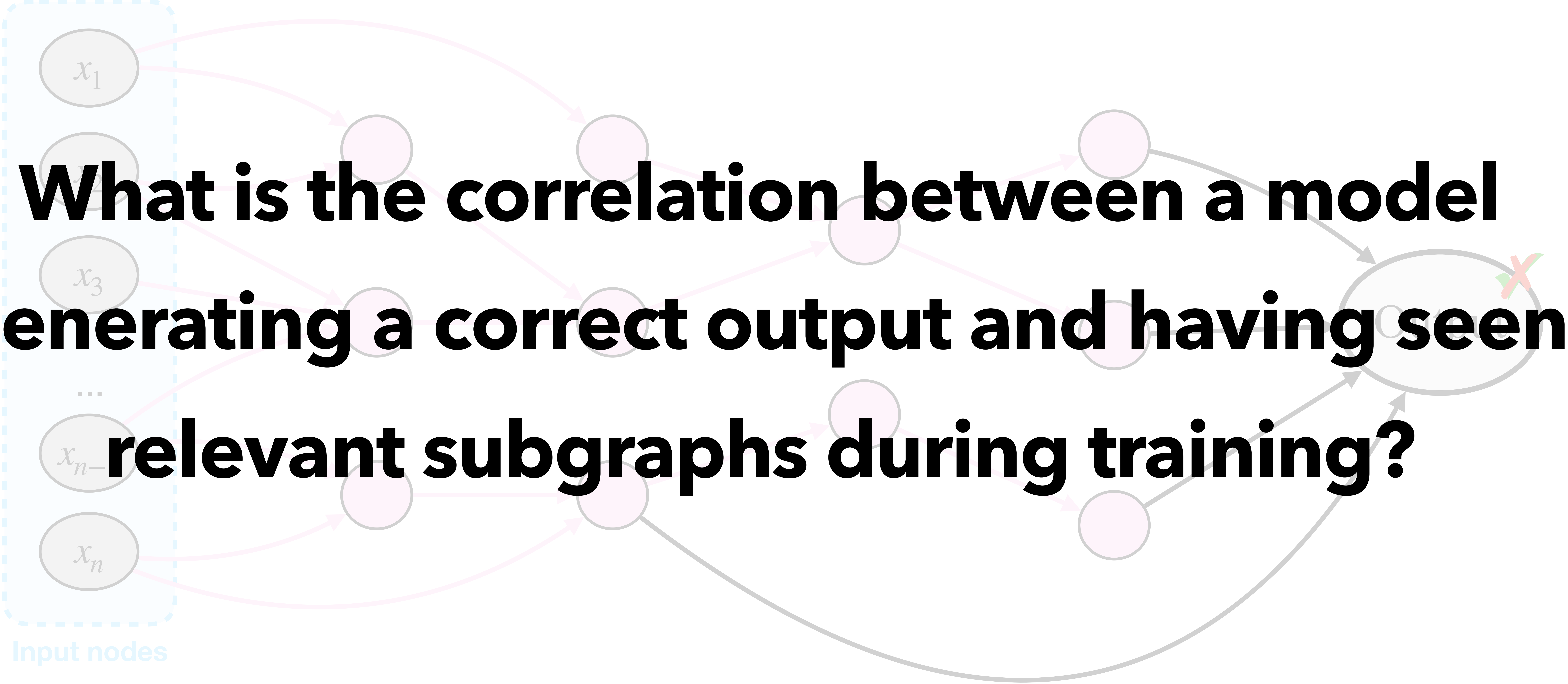
a

b =



How can we characterize model performance with respect to properties of each task sample?

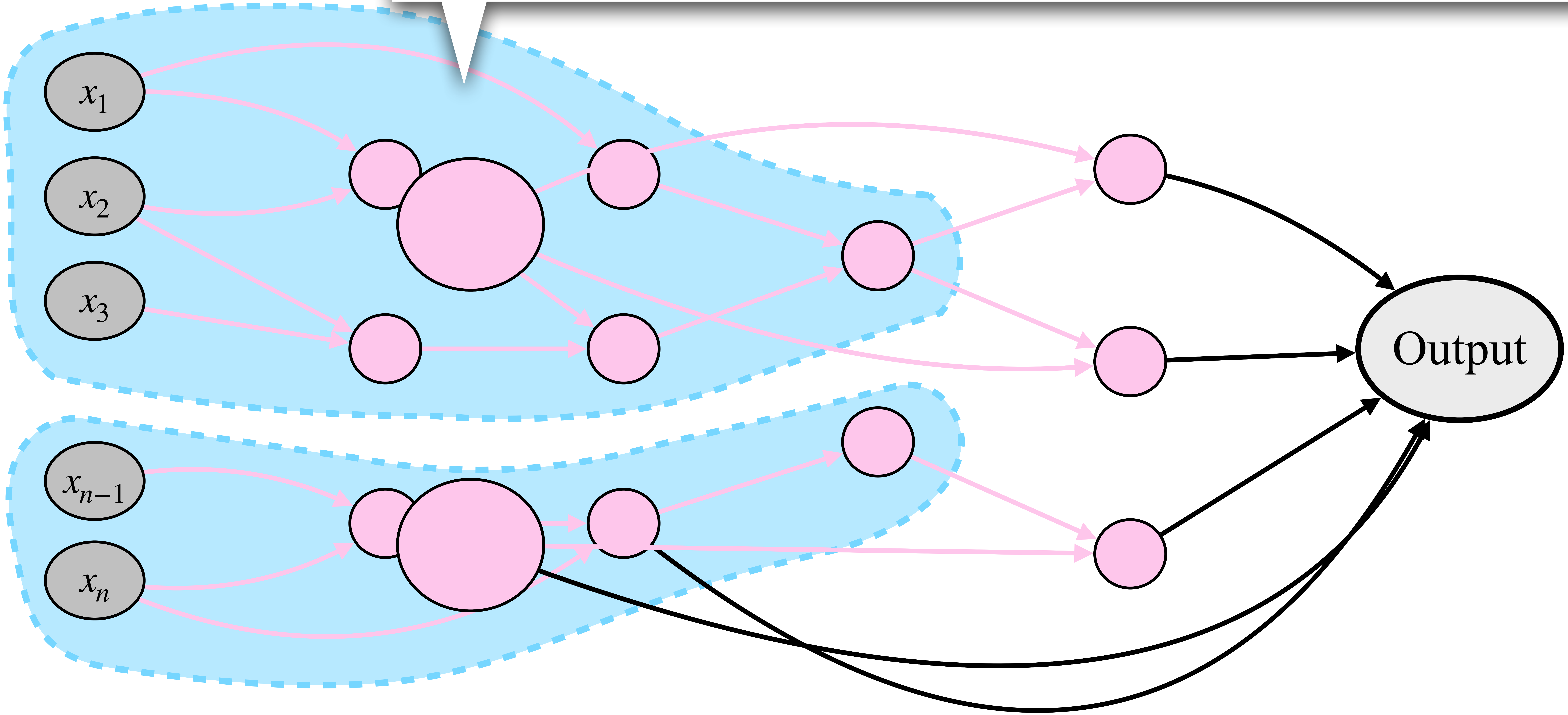
When we see models solving a seemingly difficult question, what can we correlate it to? How do we characterize model errors?



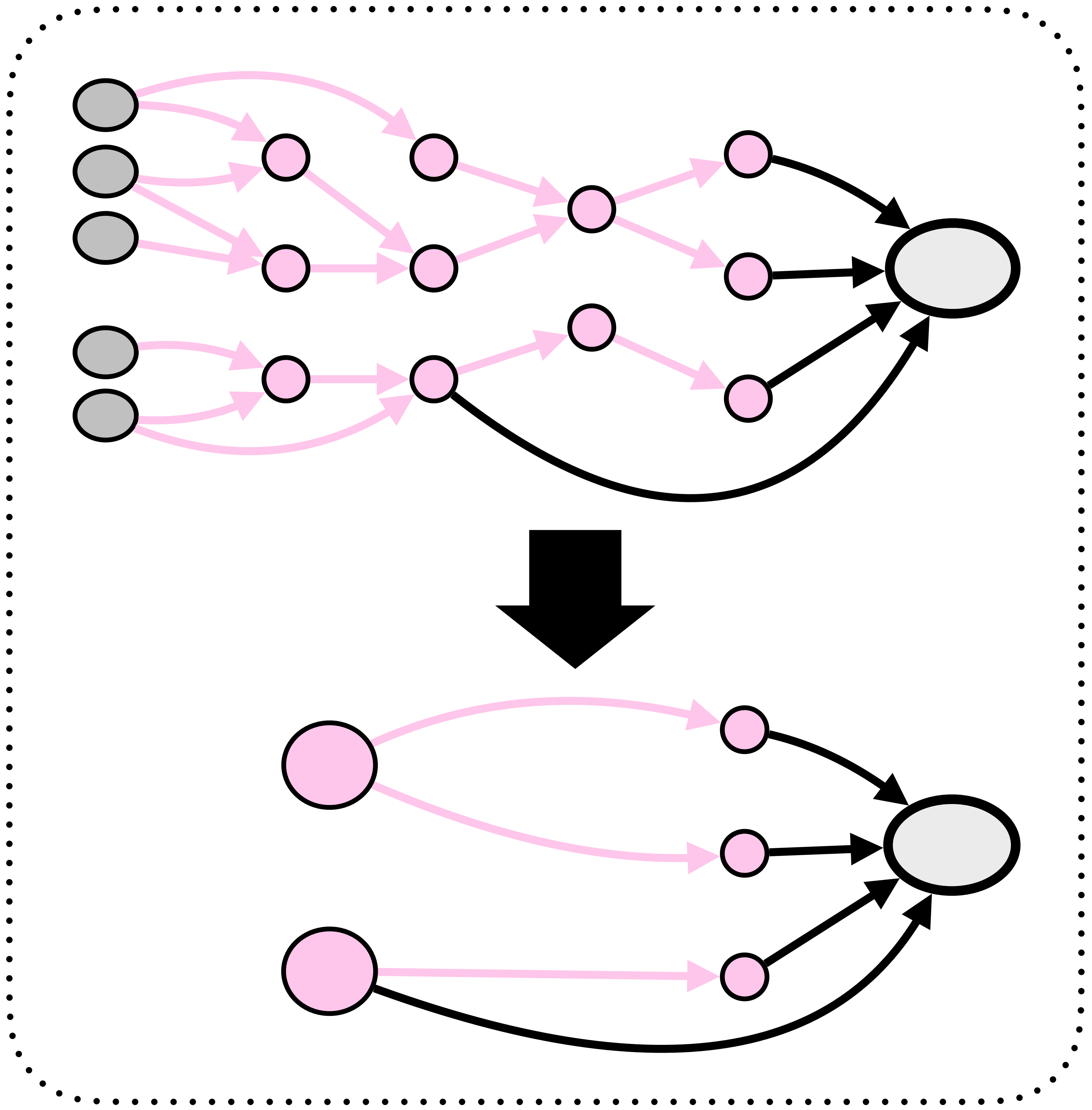
What is the correlation between a model generating a correct output and having seen relevant subgraphs during training?

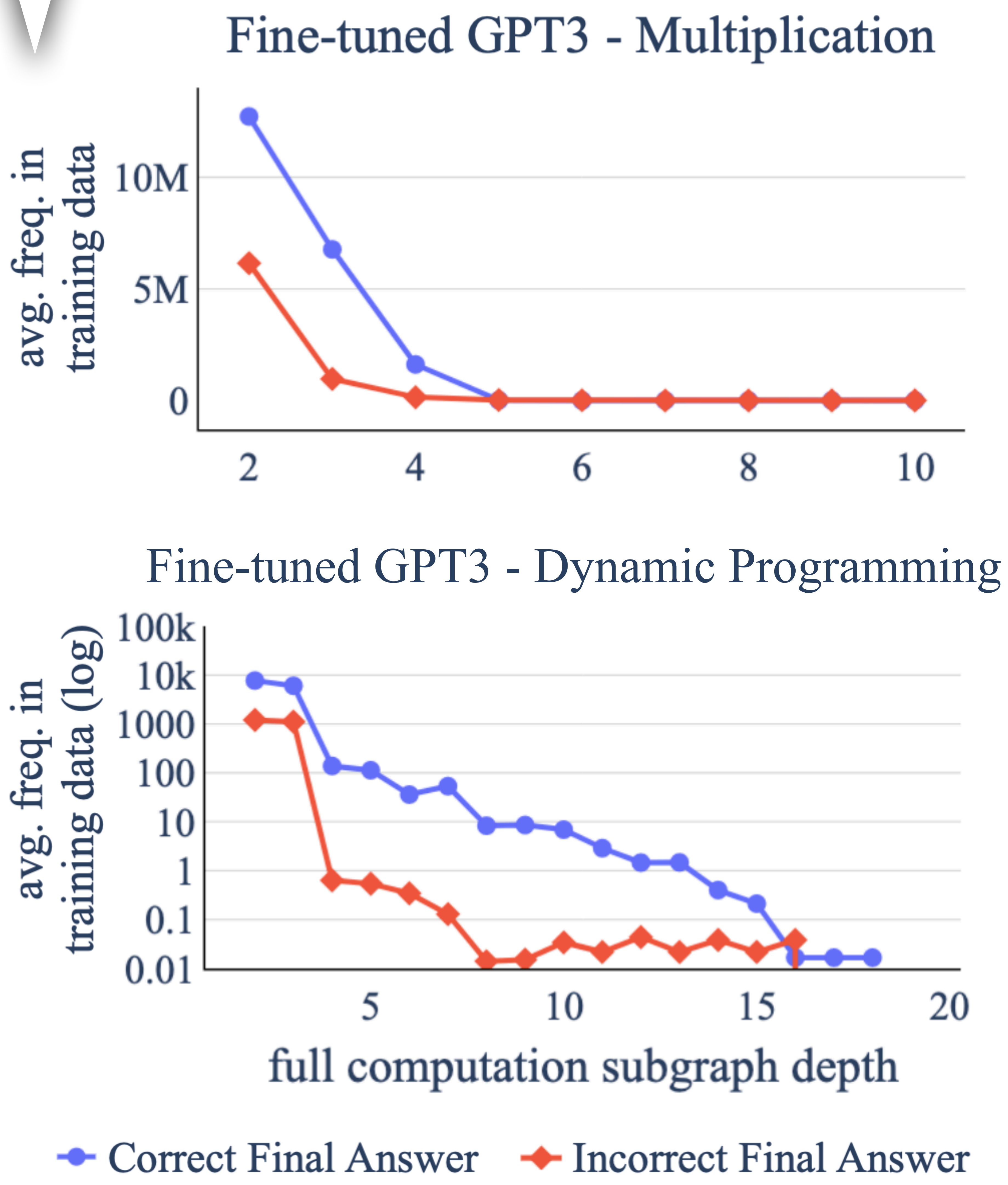
Detect subgraphs already seen during training:

vant subgraphs during training, the
inference is only *seemingly* highly compositional



Transformers' successes are heavily linked to having seen significant portions of the required computation graph during training

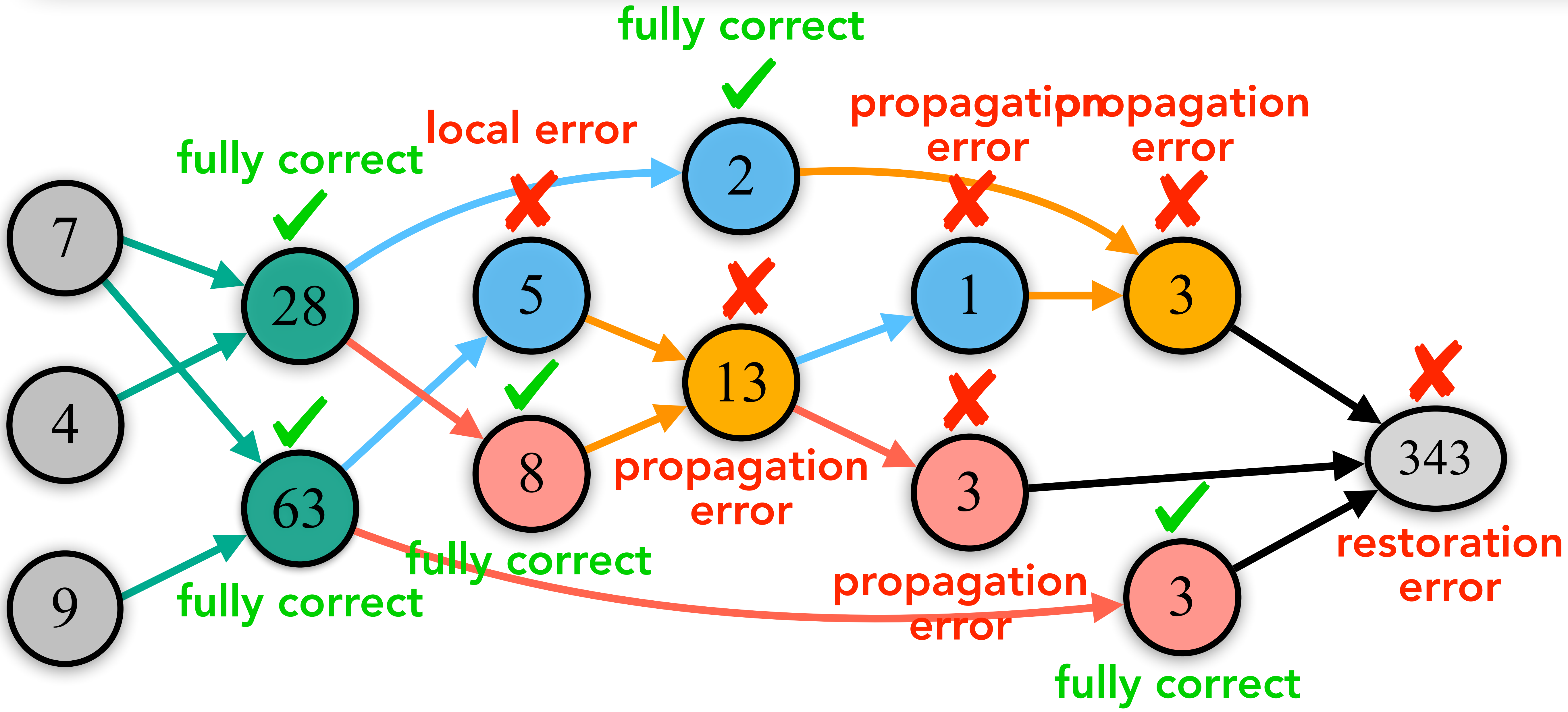


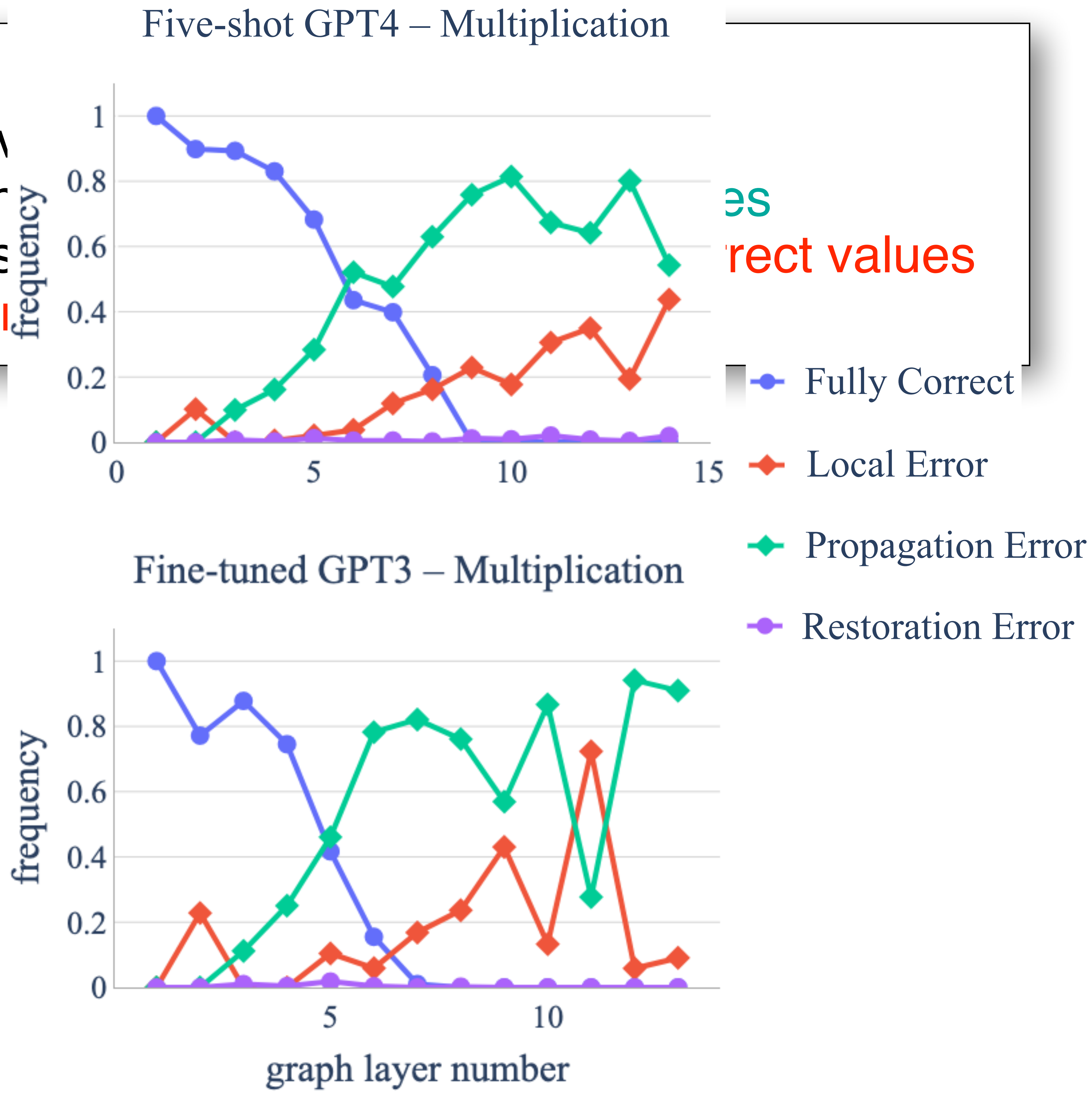


What Types of Errors do Transformers Make at Different Reasoning Depths?

Error Type

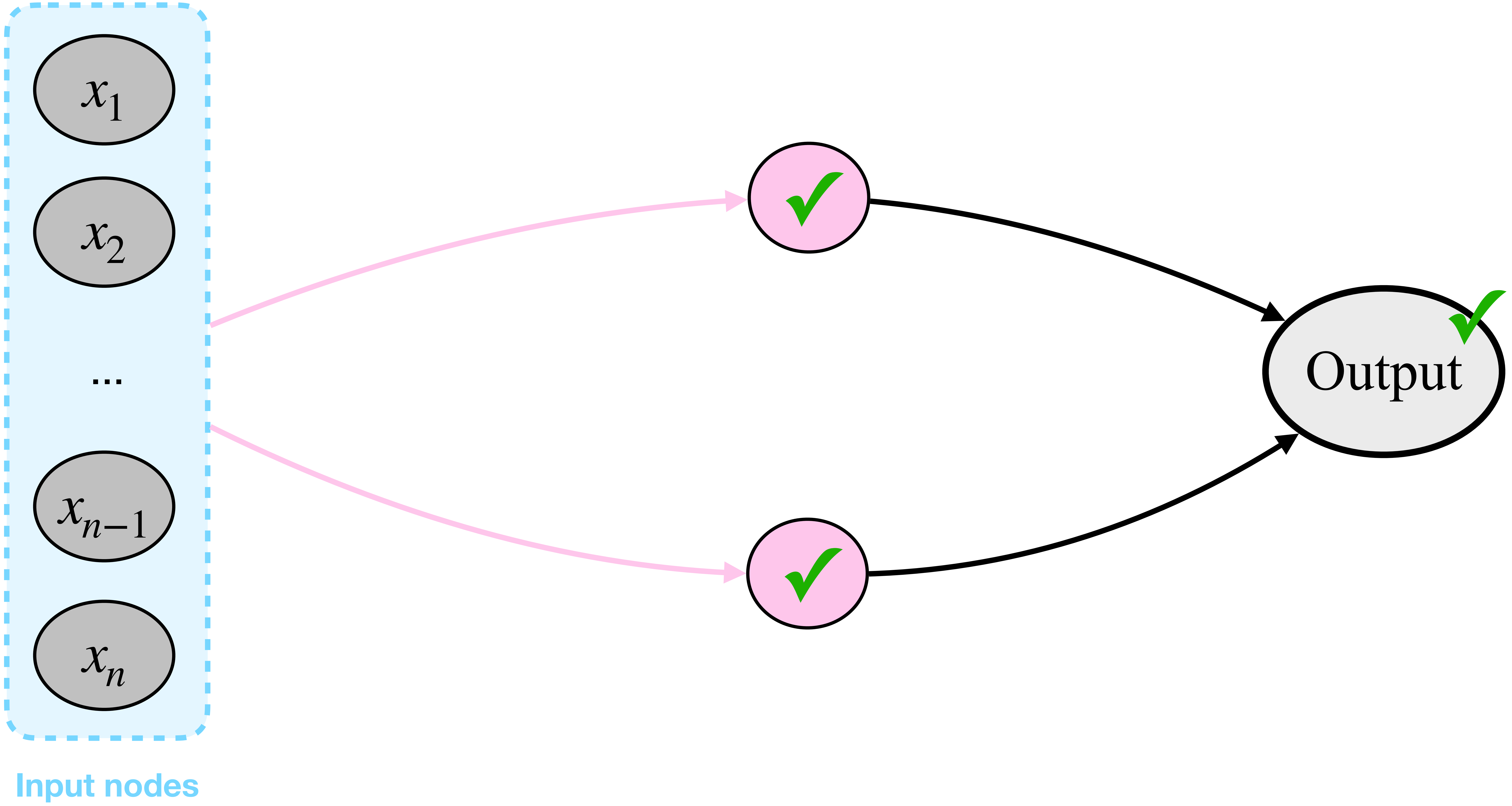
Fully Correct: v and ancestors have correct values and are deriv
Local Error: v is derived from an incorrect computation but its ar es
Propagation Error: v is derived from a correct computation but s rect values
Restoration Error: v has a correct value but is derived from an i



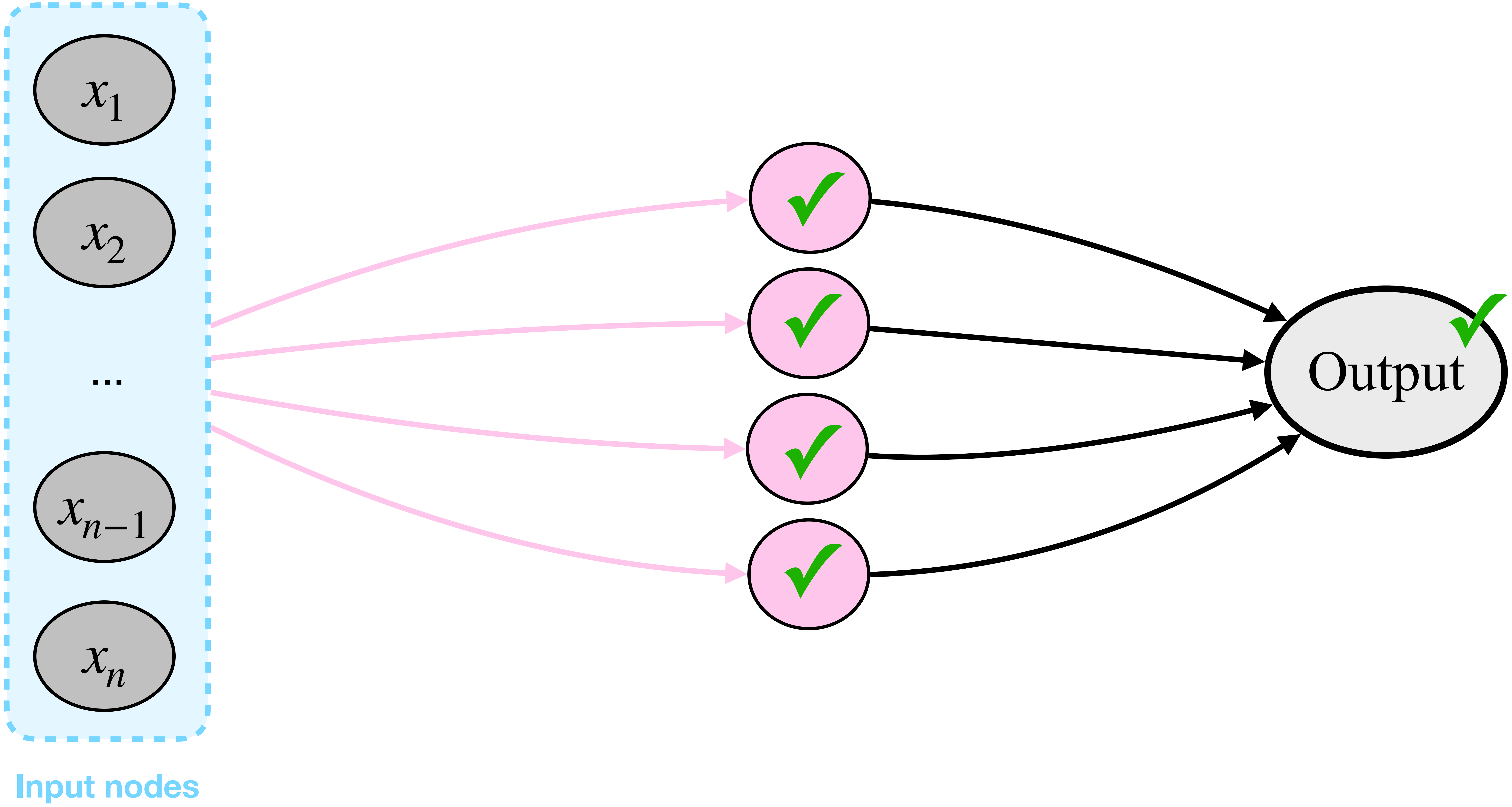


Why does performance inevitably decays with problem size?

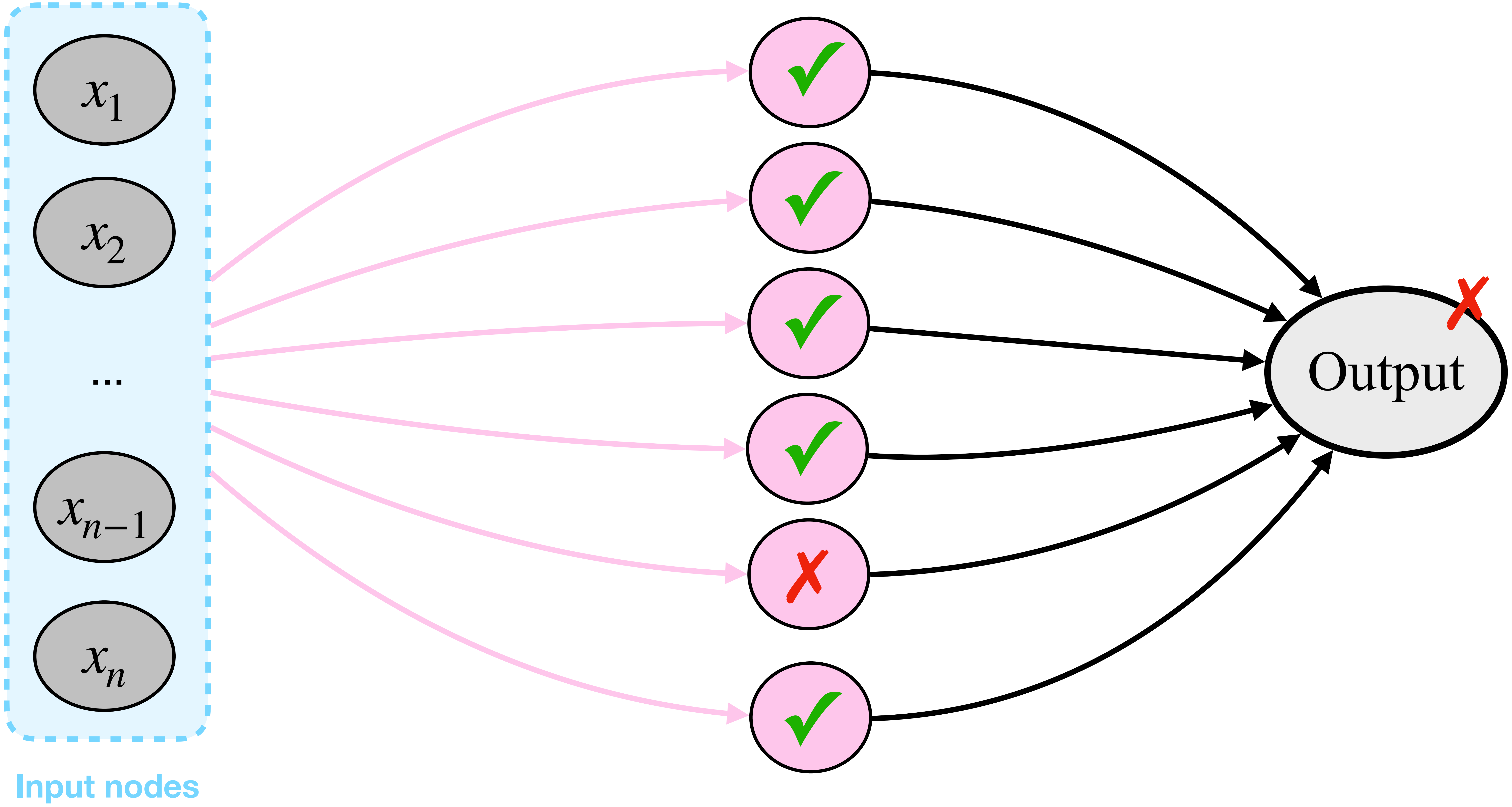
An increasing number of independent reasoning steps inevitably lead to errors



An increasing number of **independent** reasoning steps inevitably lead to errors



An increasing number of **independent** reasoning steps inevitably lead to errors



Compositional tasks often follow this pattern:

```
          9 9 9
        x 8 6 7
    -------------
        6 9 9 3
+     5 9 9 4
    7 9 9 2
    -------------
    8 6 6 1 3 3
```

Compositional tasks often follow this pattern:

$$
\begin{array}{ccccccc}
 & & & & 9 & 9 & 9 \\
 & & & \times & 8 & 6 & 7 \\
\hline
 & & & 6 & 9 & 9 & 3 \\
+ & & 5 & 9 & 9 & 4 & \\
 & 7 & 9 & 9 & 2 & & \\
\hline
 & 8 & 6 & 6 & 1 & 3 & 3
\end{array}
$$

Compositional tasks often follow this pattern:

```
          9 9 9
        x 8 6 7
    -----------
        6 9 9 3
+     5 9 9 4
    7 9 9 2
    -----------
    8 6 6 1 3 3
```

Compositional tasks often follow this pattern:

$$
\begin{array}{ccccccc}
 & & & & 9 & 9 & 9 \\
 & & & \times & 8 & 6 & 7 \\
\hline
 & & & 6 & 9 & 9 & 3 \\
+ & & 5 & 9 & 9 & 4 & \\
 & 7 & 9 & 9 & 2 & & \\
\hline
 & 8 & 6 & 6 & 1 & 3 & 3
\end{array}
$$

Compositional tasks often follow this pattern:

$$
\begin{array}{rcccccc}
 & & & & 9 & 9 & 9 \\
 & & & \times & 8 & 6 & 7 \\
\hline
 & & & 6 & 9 & 9 & 3 \\
+ & & 5 & 9 & 9 & 4 & \\
 & 7 & 9 & 9 & 2 & & \\
\hline
 & 8 & 6 & 6 & 1 & 3 & 3
\end{array}
$$

Compositional tasks often follow this pattern:

$$
\begin{array}{ccccccc}
 & & & & 9 & 9 & 9 \\
 & & & \times & 8 & 6 & 7 \\
\hline
 & & & 6 & 9 & 9 & 3 \\
+ & & 5 & 9 & 9 & 4 & \\
 & 7 & 9 & 9 & 2 & & \\
\hline
 & 8 & 6 & 6 & 1 & 3 & 3
\end{array}
$$

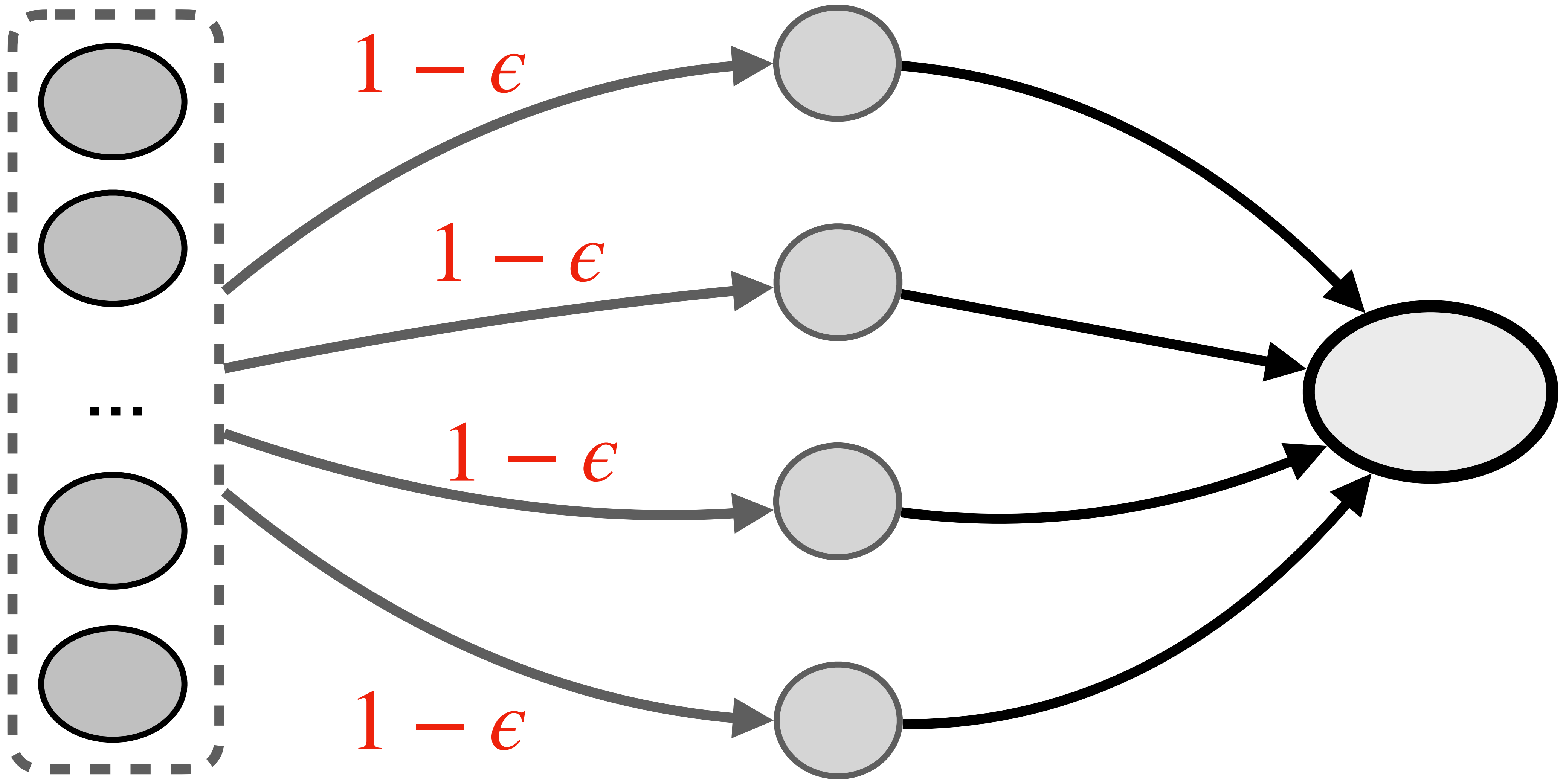
Compositional tasks often follow this pattern:

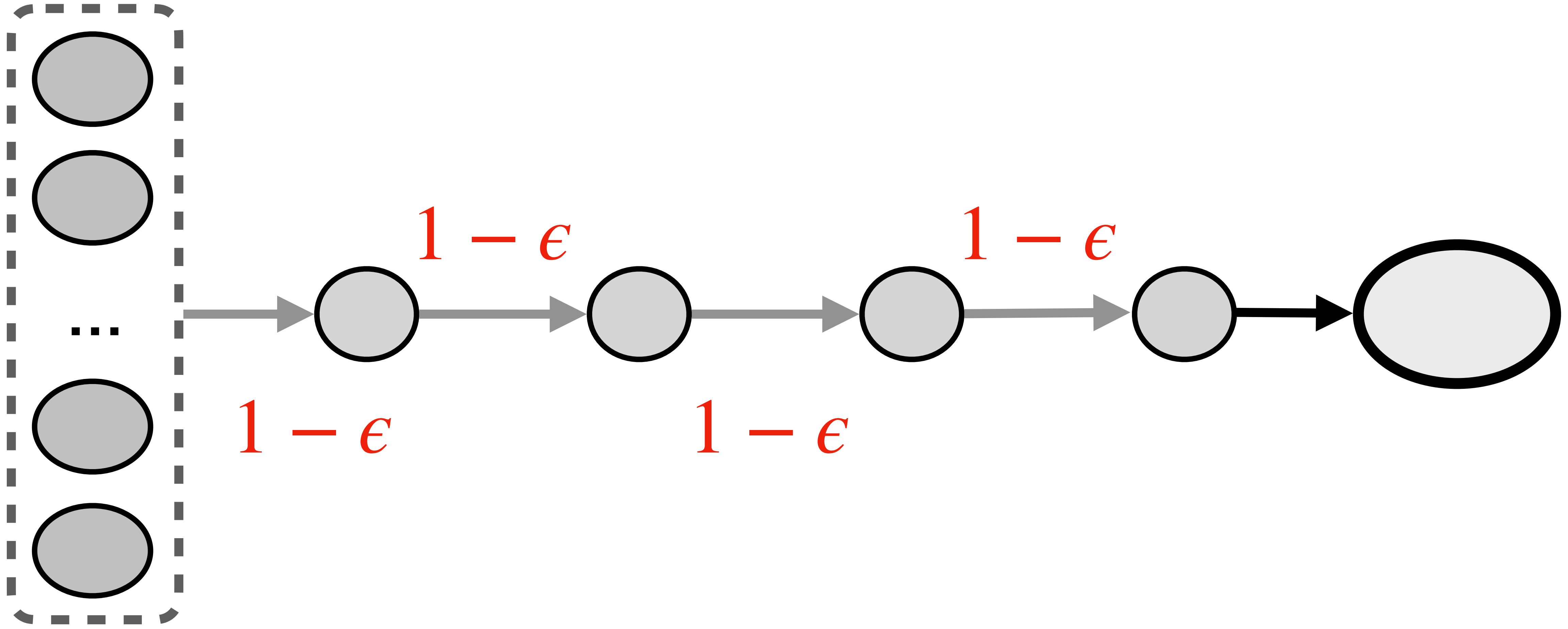
```
        9 9 9
      x 8 6 7
   -----------
      6 9 9 3
+   5 9 9 4
  7 9 9 2
   -----------
  8 6 6 1 3 3
```

An increasing number of reasoning steps inevitably ~~lead to~~ errors

with exponentially increasing probability

If the probability of making an error in a single reasoning step is ϵ, probability of success is…





$$\approx (1 - \epsilon)^n$$

$$\approx (1 - \epsilon)^n$$

Theoretical framing on error accumulation may inform future developments

- Does it mean there aren't avenues for progress? No! Promising avenues:
  - use transformers in ways that **chain only few compositional steps** to reach a solution
  - use transformers in tasks where **evaluation metrics afford leniency**
  - **augmenting transformers with planning modules** and refining methods to decrease ϵ!

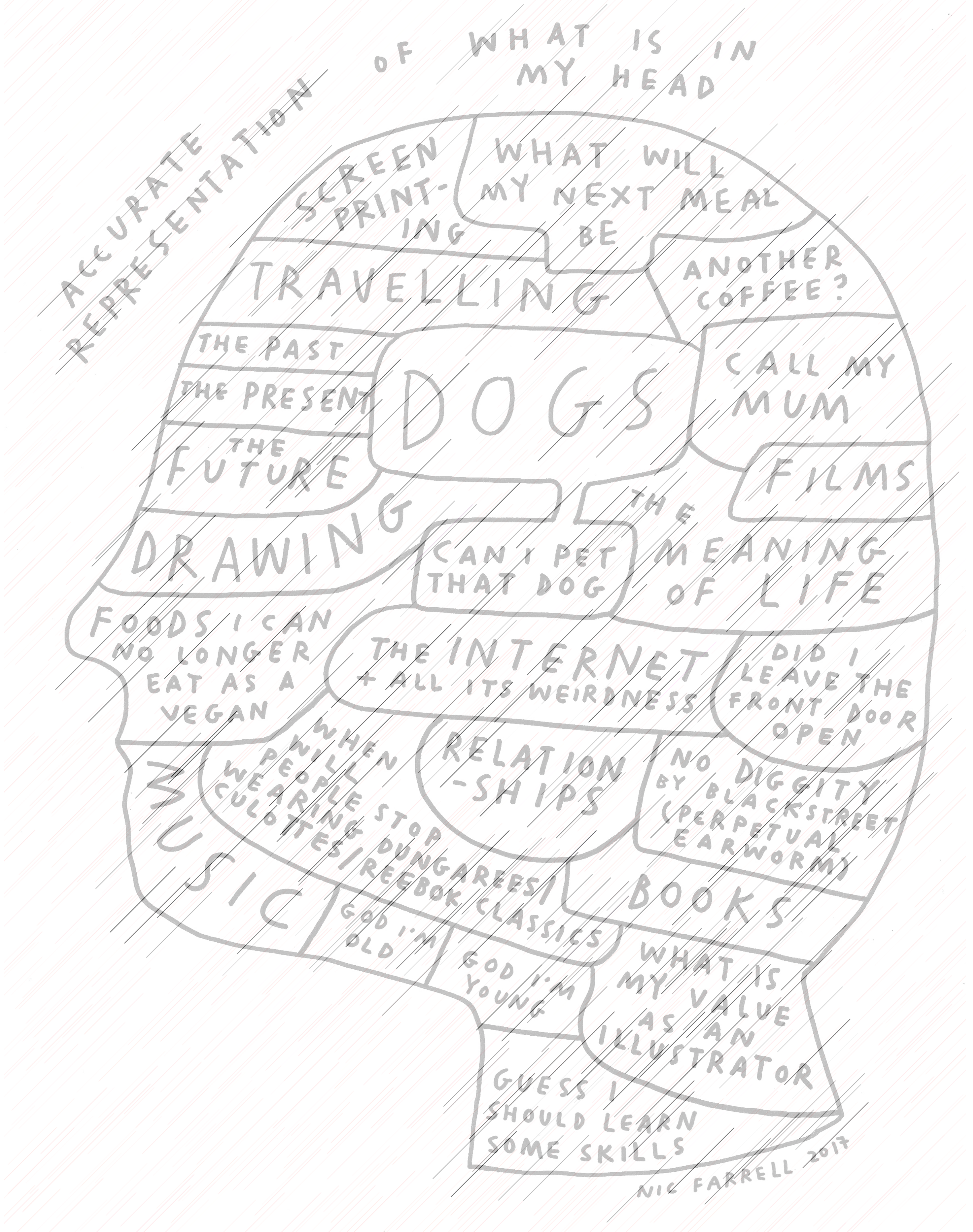
Faithful Reasoning Evaluation

Complex ***conclusion-based*** *evaluation for theory of mind reasoning*

Theory of Mind

the ability to reason about the mental states **of others**

e.g., desires, beliefs, intentions, etc.

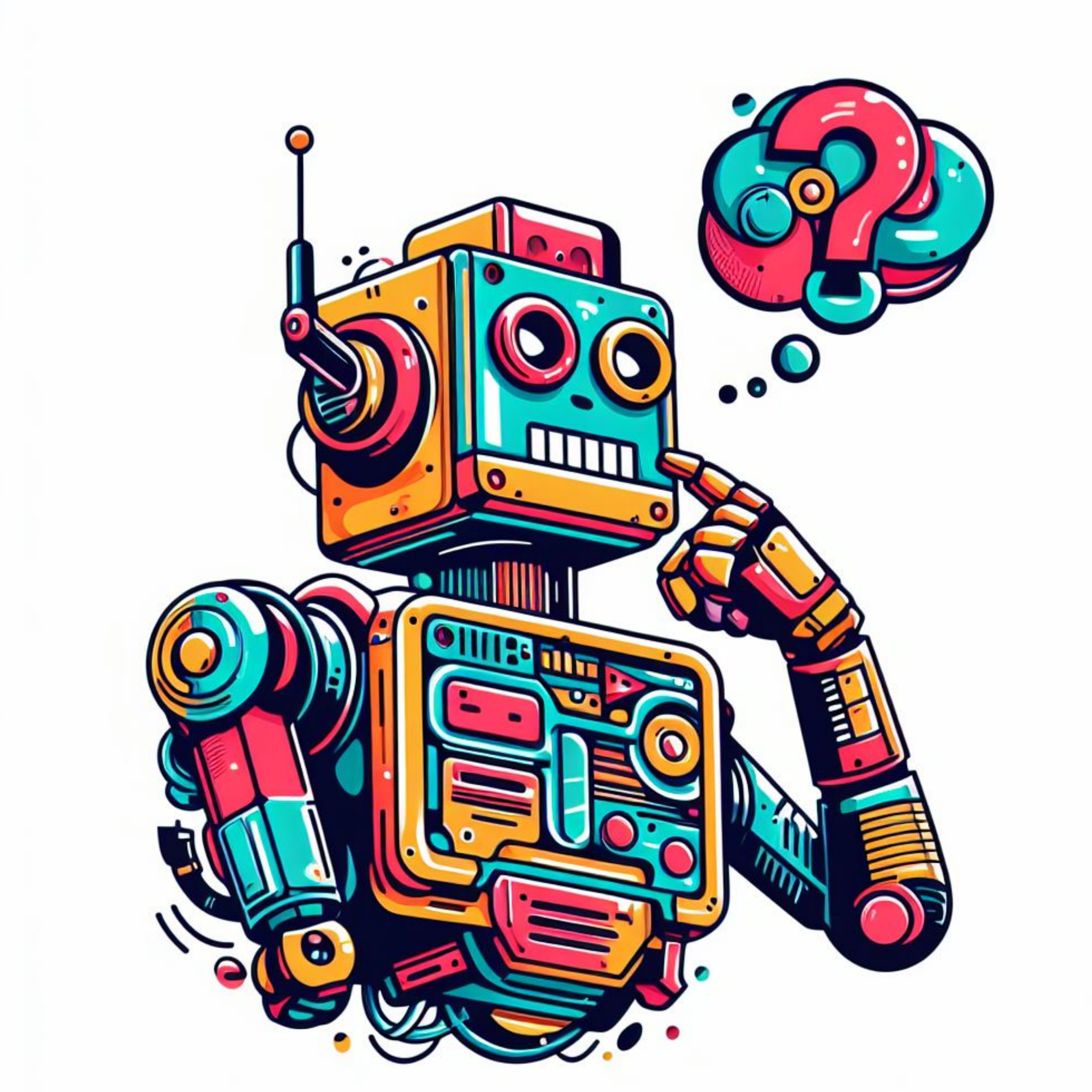


Theory of Mind?

Are we saying machines have a mind?

No, they do not have minds, emotions, or intentions

However, they need social reasoning capabilities



What is theory of mind/social cognition?

One of the most quintessential human mental function:

Thinking about each other's thoughts

- Our relationship with other people is the most crucial aspect of our lives
- Social cognition takes up a huge part of our reasoning
  - Every minute! Even right now
  - Social factors impacted the evolution of our intelligence

GPT-4 already shows sparks of AGI?

**Sparks of Artificial General Intelligence:
Early experiments with GPT-4**

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Microsoft Research

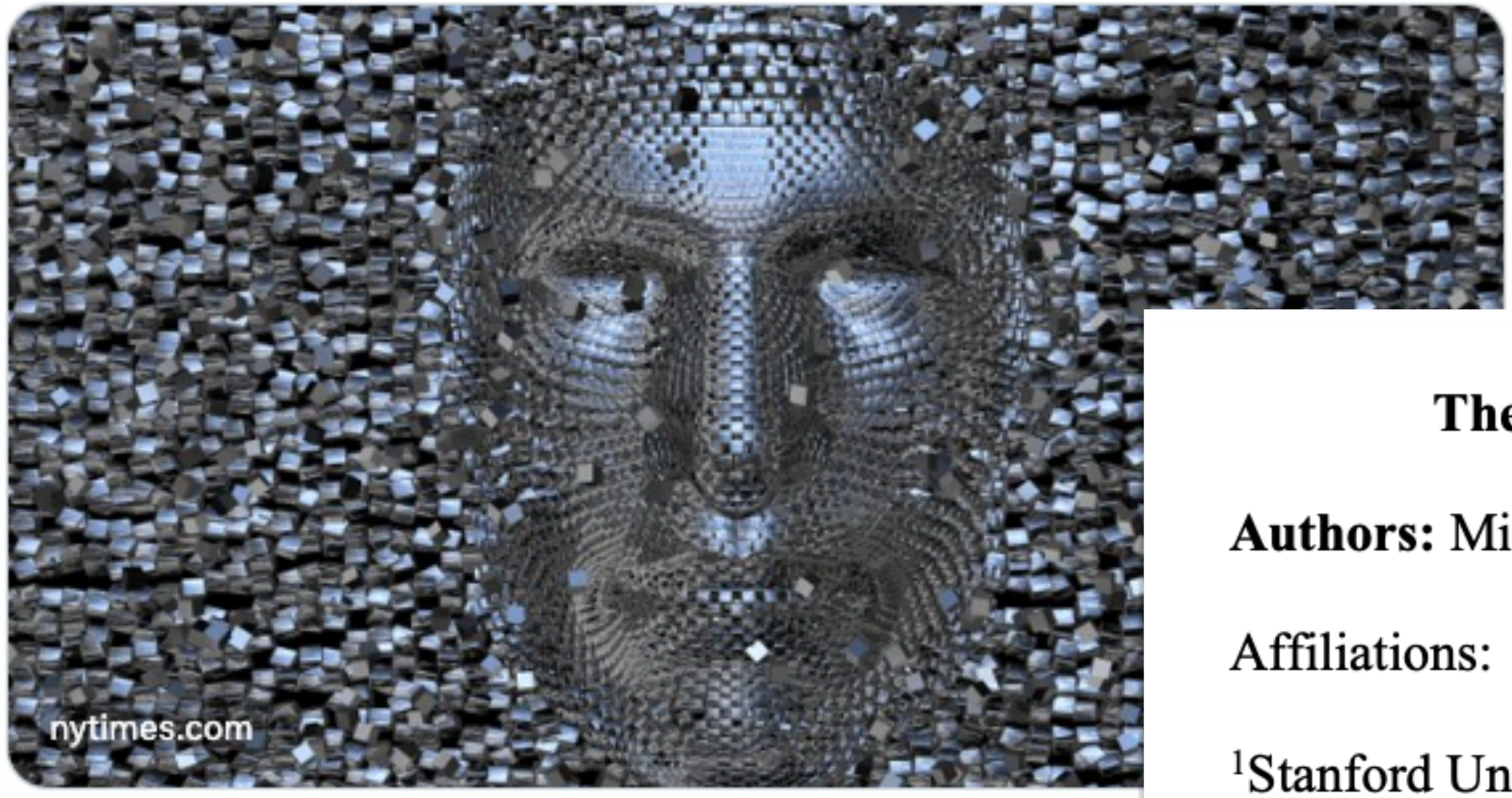
"Our findings suggest that ***GPT-4 has a very advanced level of theory of mind.****"*

from 6 examples

GPT-4 already has theory of mind?

Stanford Graduate School of Business
@StanfordGSB

Machines have theory of mind.
That's the conclusion Professor @michalkosinski reached in his latest research. Using #ChatGPT, Kosinski has been testing what mental capacities these #AI tools may have. @nytimes



5:00 PM · Apr 6, 2023 · **15.8K** Views

7 22 77 17

Theory of Mind Might Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305, USA

GPT-4 already has theory of mind?

ChatGPT 4

This is the "Smarties task", a famous test in psychology

HY **You**

There is a closed bag filled with chocolate. The label on the bag says "popcorn".
Sam finds the bag. She has seen the bag for the first time and did not look inside.

Sam will think that she has just found a bag full of

ChatGPT

popcorn.

Theory of Mind Might Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305, USA

Look again

Even a slightest change in the label with an unseen word causes GPT-4 to fail on these theory of mind tests.

ChatGPT 4

You

There is a closed bag filled with chocolate. The label on the bag say "UWISGREAT". Sam saw this bag for the first time and did not look inside.

Sam thinks that she has just found a bag full of

ChatGPT

Based on your description, Sam might think that she has found a bag full of chocolate, as indicated by the bag's appearance and label. Would you like me to create an image of this scenario?

Look again

ChatGPT 4

You

There is a closed white bag filled with chocolate. The label on the bag say "UWISGREAT". Sam saw this bag for the first time and did not look inside.

Sam thinks that she has just found a bag full of what?

ChatGPT

Based on your description, Sam, upon seeing the closed white bag labeled "UWISGREAT" filled with chocolate but without looking inside, would likely assume that she has found a bag full of chocolates or sweets. This assumption is based on the context that the bag is specifically filled with chocolate, and the label does not provide any contradictory information to suggest otherwise.

Look again

ChatGPT 4

You

There is a closed brown bag filled with chocolate. The label on the bag just say "UWISGREAT" without any details. Sam saw this bag for the first time and did not look inside.

Sam thinks that she has just found a bag full of

ChatGPT

UWISCHOCOLATE

Development of ToM

Recognize that others have

1. Diverse desires
2. Diverse beliefs soon after
3. Access to different knowledge bases
4. May have False beliefs
5. Capability of hiding emotions

Development of ToM

Recognize that others have

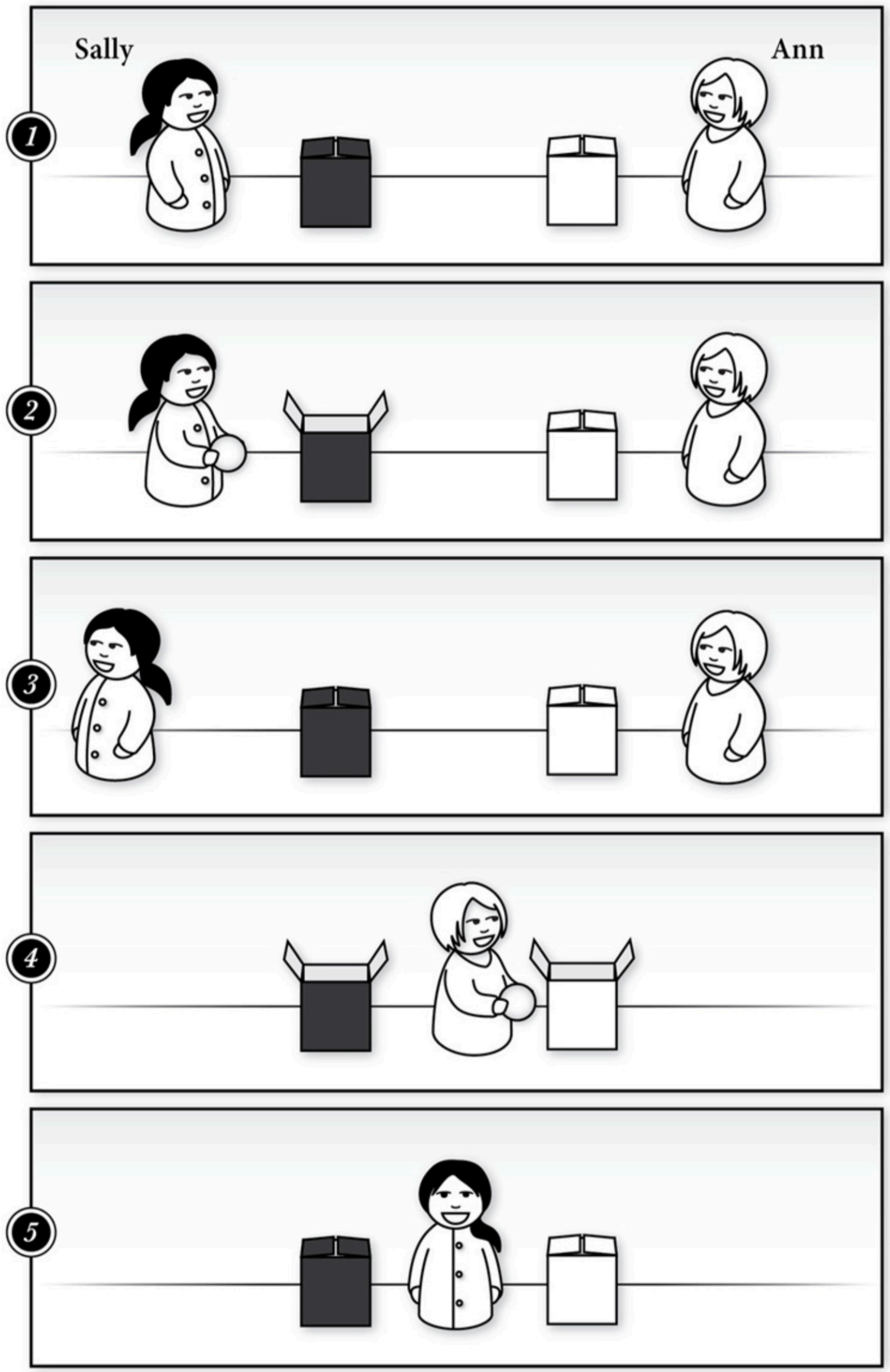
1. Diverse desires
2. Diverse beliefs soon after
3. Access to different knowledge bases
4. **May have False beliefs**
5. Capability of hiding emotions

The Sally-Anne test

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

1. Sally has a black box and Anne has a white box.
2. Sally has a marble. She puts the marble into her box.
3. Sally goes for a walk.
4. Anne takes the marble out of Sally's box and puts into her box.
5. Sally comes back and wants to play with her marble.

Question: Where will Sally look for her marble?



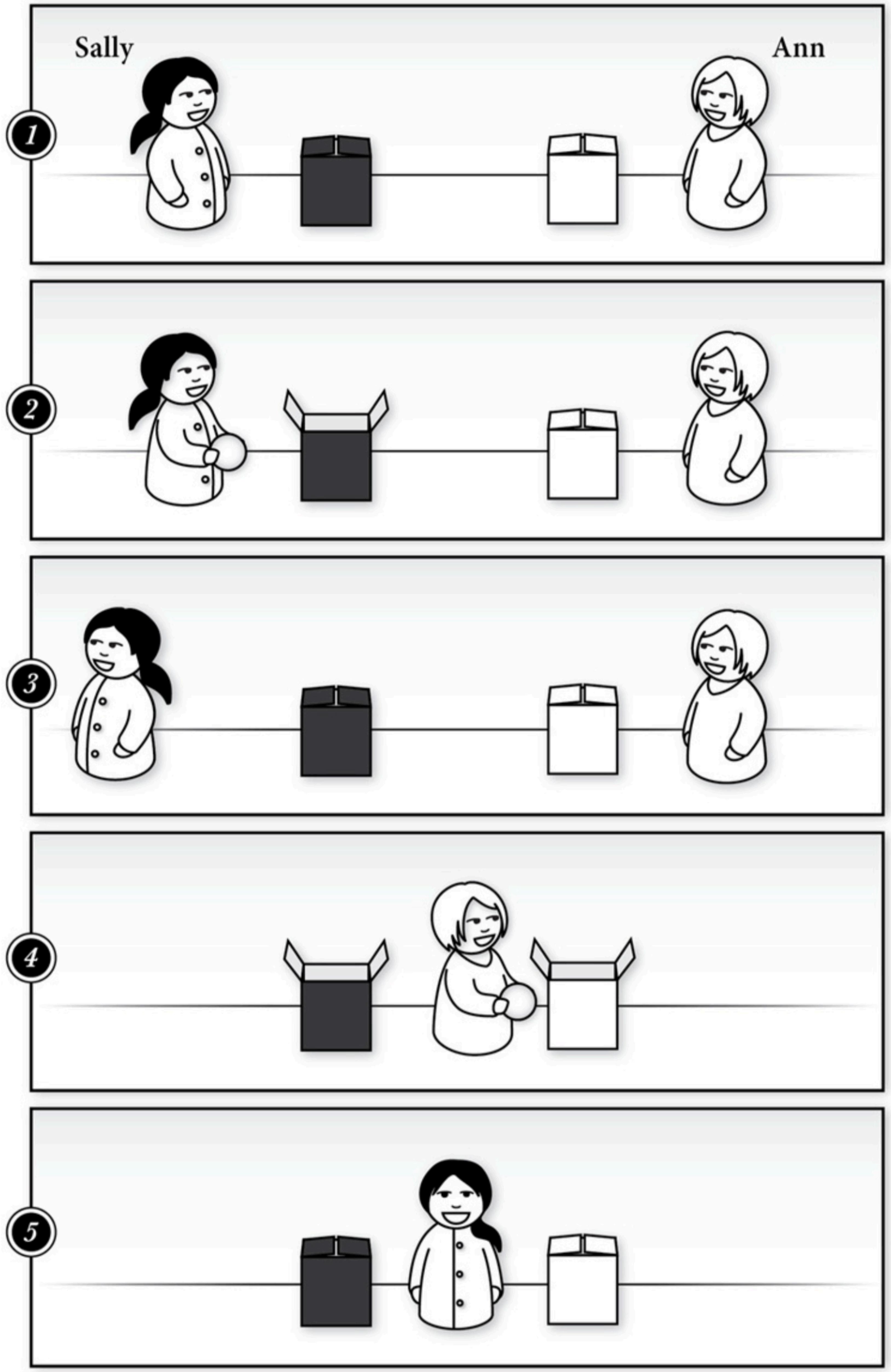
The Sally-Anne test

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Question: Where will Sally look for her marble?

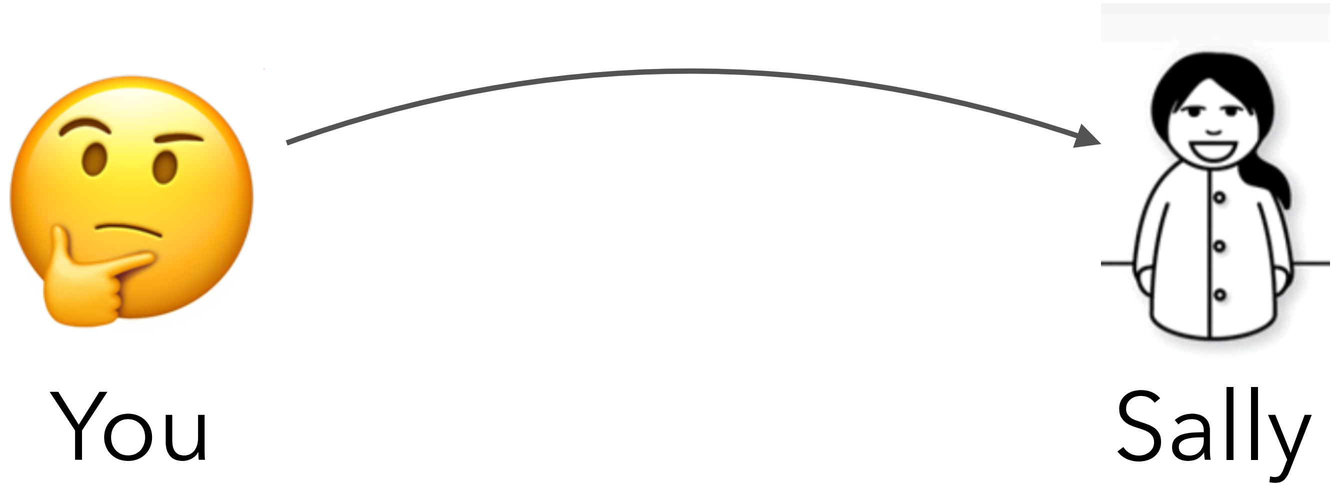
- Before the age of 4: Sally will look for it in Anne's box
- By the age of 4: Sally will look for it in her box

By the age of 4, children begin to understand that others may have ***false beliefs***

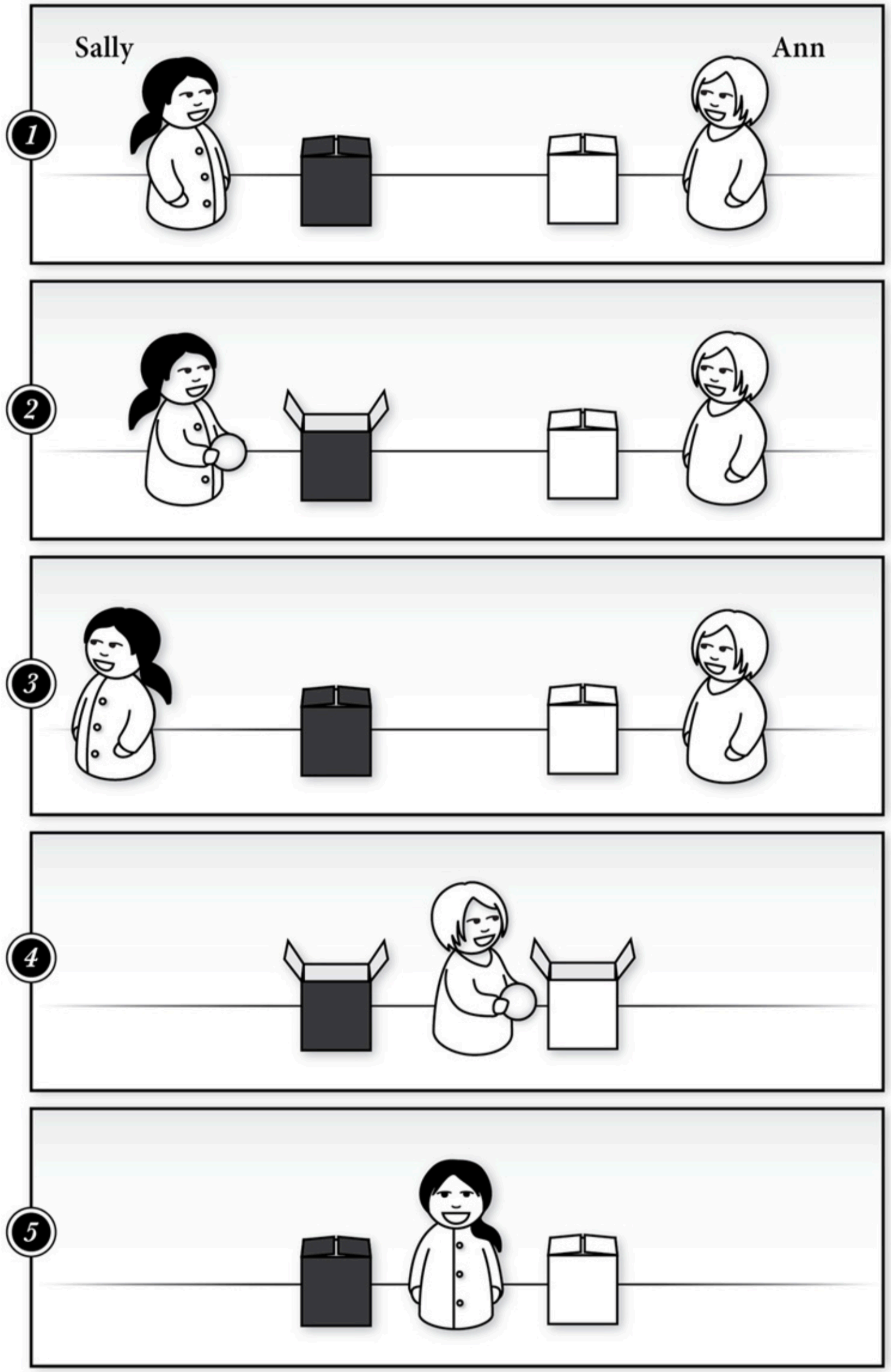


Order of ToM

Where will Sally think her marble is?

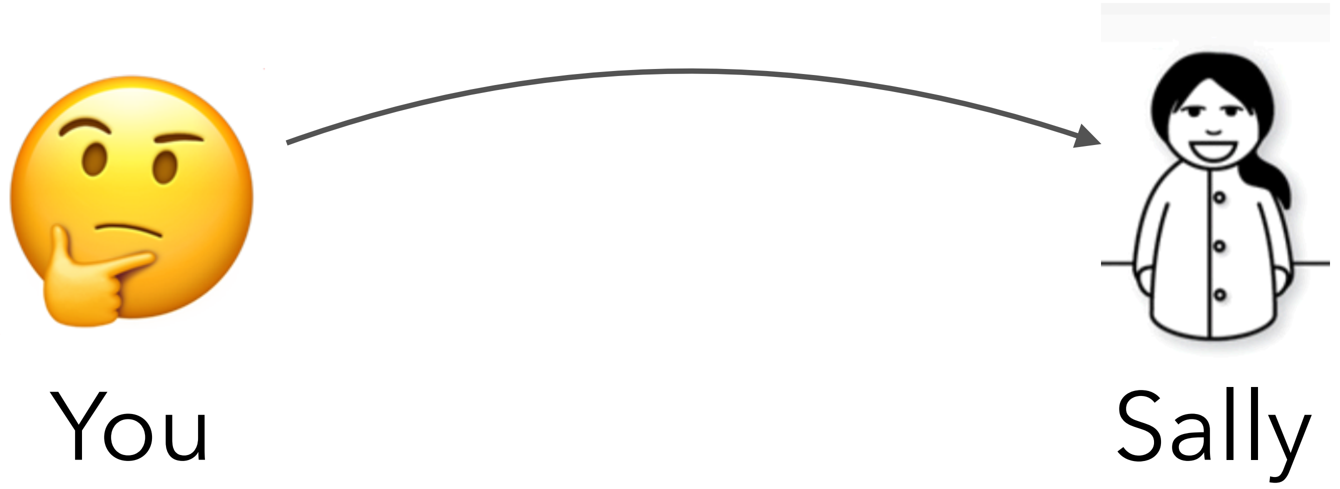


First-order



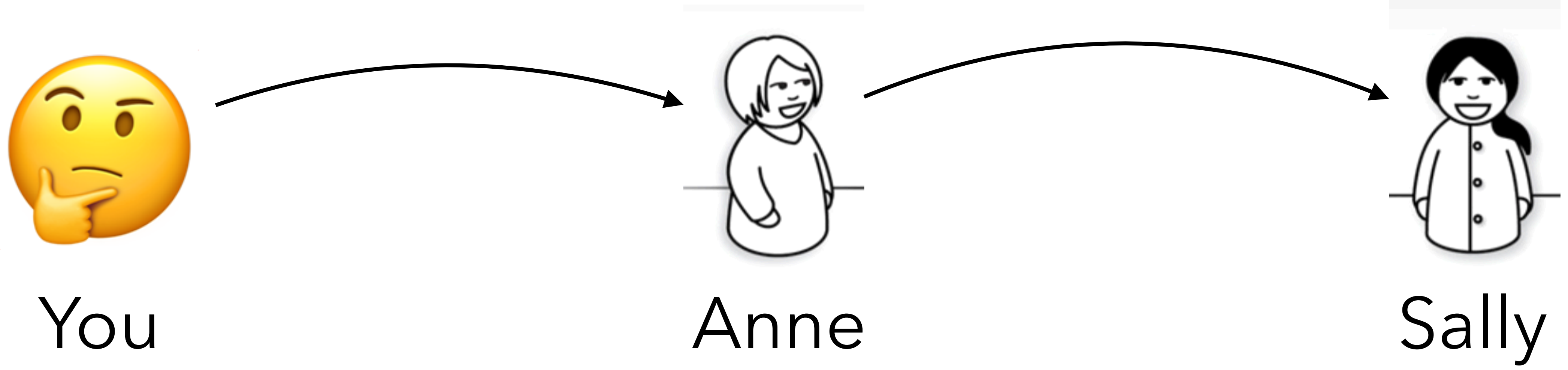
Order of ToM

Where will Sally think her marble is?

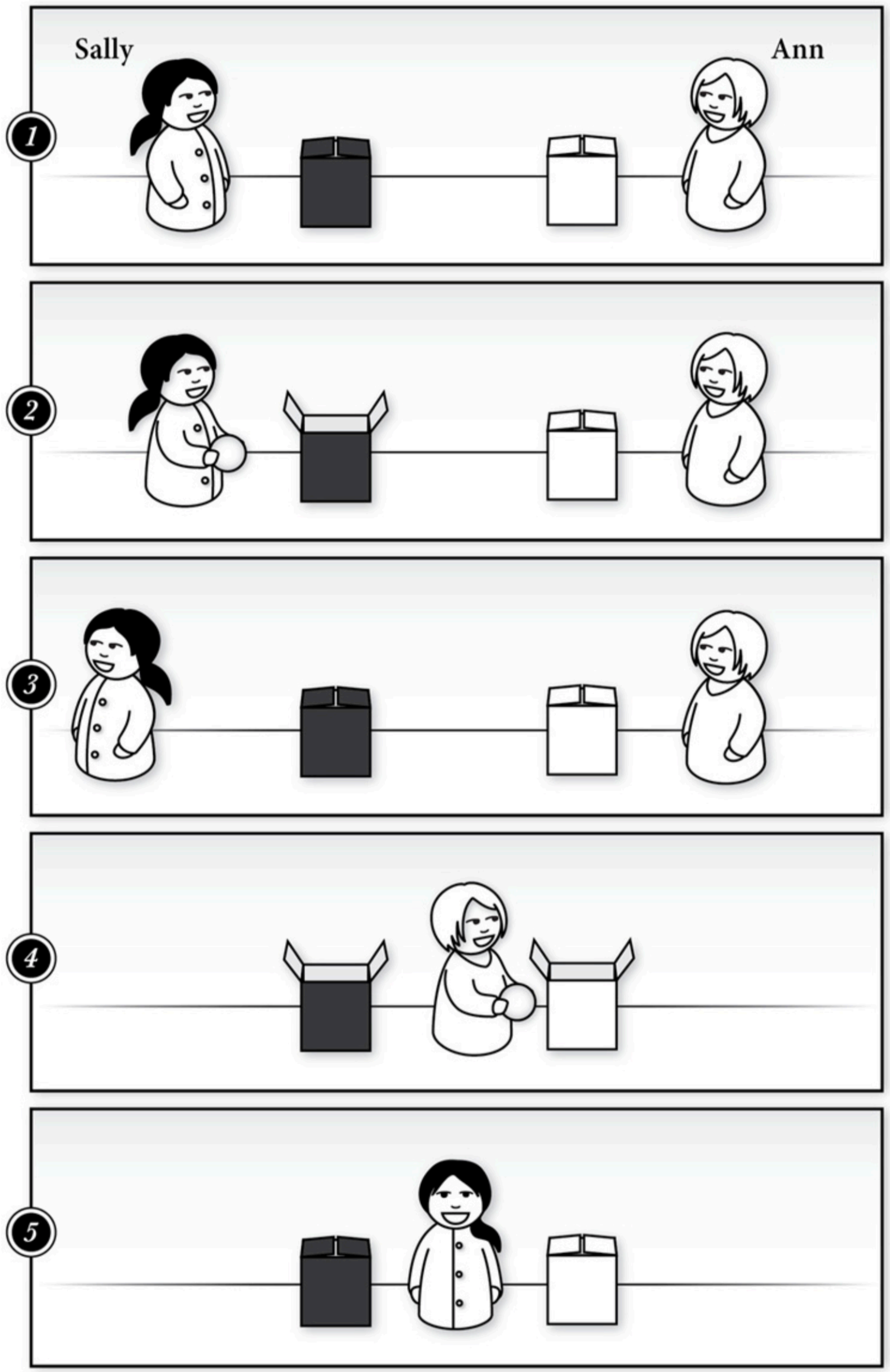


First-order

Where will Anne think
Sally thinks her marble is?



Second-order



How can we systematically quantify theory of mind reasoning skills?

TrackTheMind: program-guided adversarial data generation for theory of mind reasoning

In submission

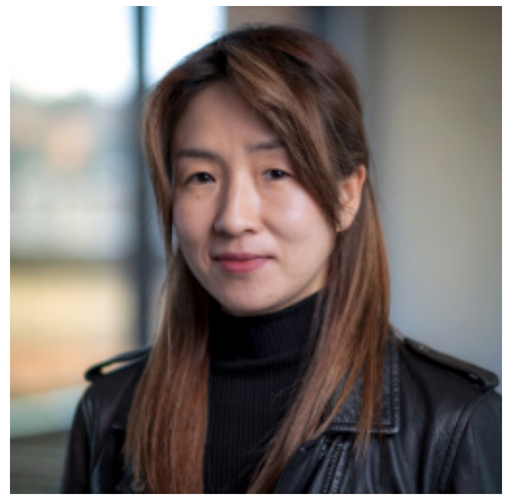

Melanie Sclar


Jane Dwivedi-Yu


Maryam Fazel-Zarandi


Yulia Tsvetkov


Yonatan Bisk


Yejin Choi


Asli Celikyilmaz

Motivation

- Theory of mind skills are difficult to measure
  - Hard to find enough explicit ToM data in the wild
  - Data leaks
  - Accidentally evaluating on easy cases (models are improving!)
- **Let's automatically generate difficult ToM data so we can stress-test models! Specifically (story, ToM question, answer) triples.**

Theory of Mind-specific domain language: capabilities

- **We code a small "world model":** we automatically track the mental state updates any time someone performs an action
- **Actions supported:**
  - entering and leaving a room,
  - moving objects to a container or another room,
  - changing the state of an object,
  - communicating with people about abstract topics or to tell them about a world state change, asymmetry (people spying or being distracted)

Theory of Mind-specific domain language: example

<story start>

Anne entered the kitchen.

update #1

worldState[Anne, location] -> kitchen
belief[Anne, location] -> kitchen

Then…

Beth entered the kitchen.

update #2

worldState[Beth, location] -> kitchen
belief[Anne, Beth, location] -> kitchen
belief[Beth, Anne, location] -> kitchen

Then…

Beth salted the apple.

update #3

belief[Anne, apple, salted] -> yes
belief[Anne, Beth, apple, salted] -> yes
…

…

Beth left the kitchen.

Charles moved the apple to the fridge.

Charles entered the kitchen.

Beth texted to Charles to let him know the apple is salted.

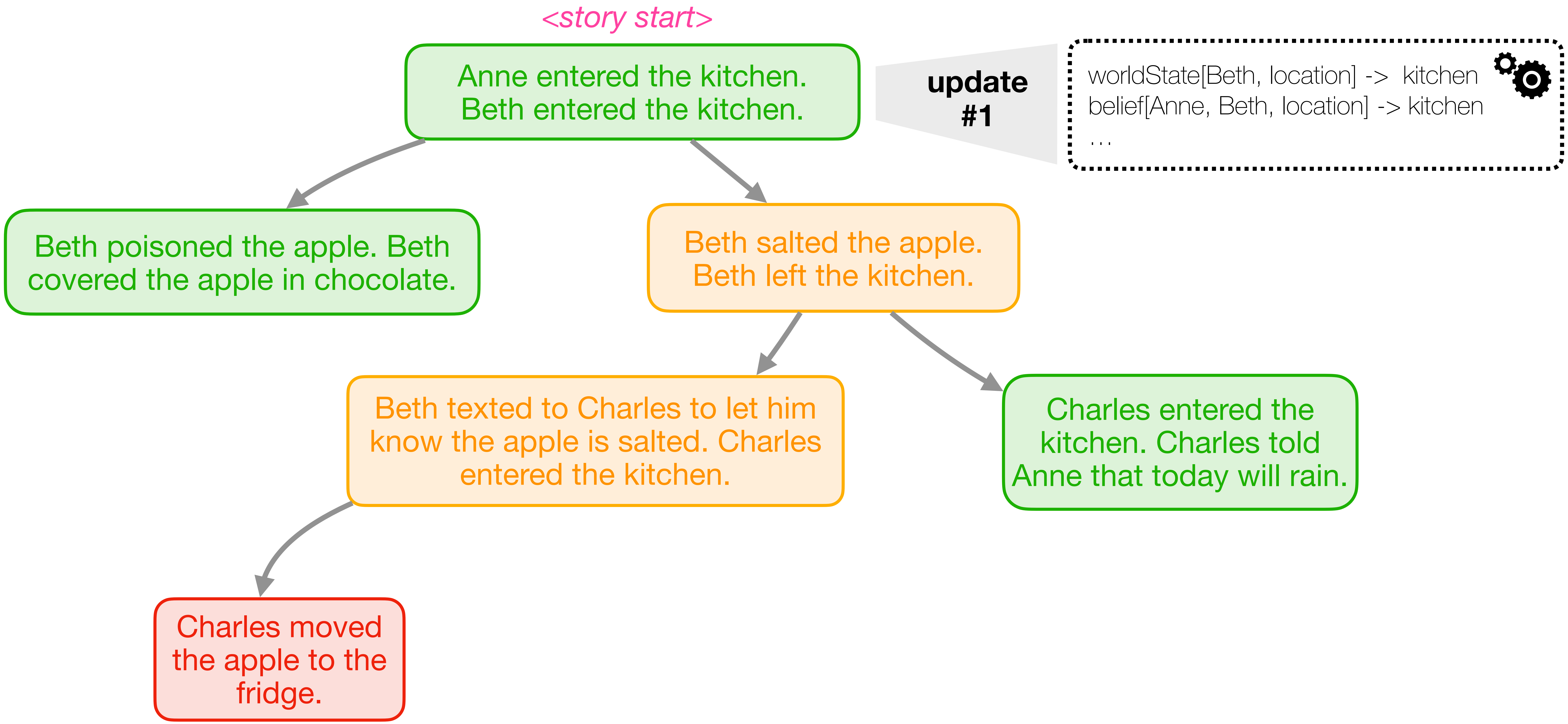
tracker-generated questions

Where is the apple right now? ***fridge***
Where does Beth think the apple is? ***table***
Does Anne know that the apple is salted? ***yes***
Does Anne think Charles knows that the apple is salted? ***no***

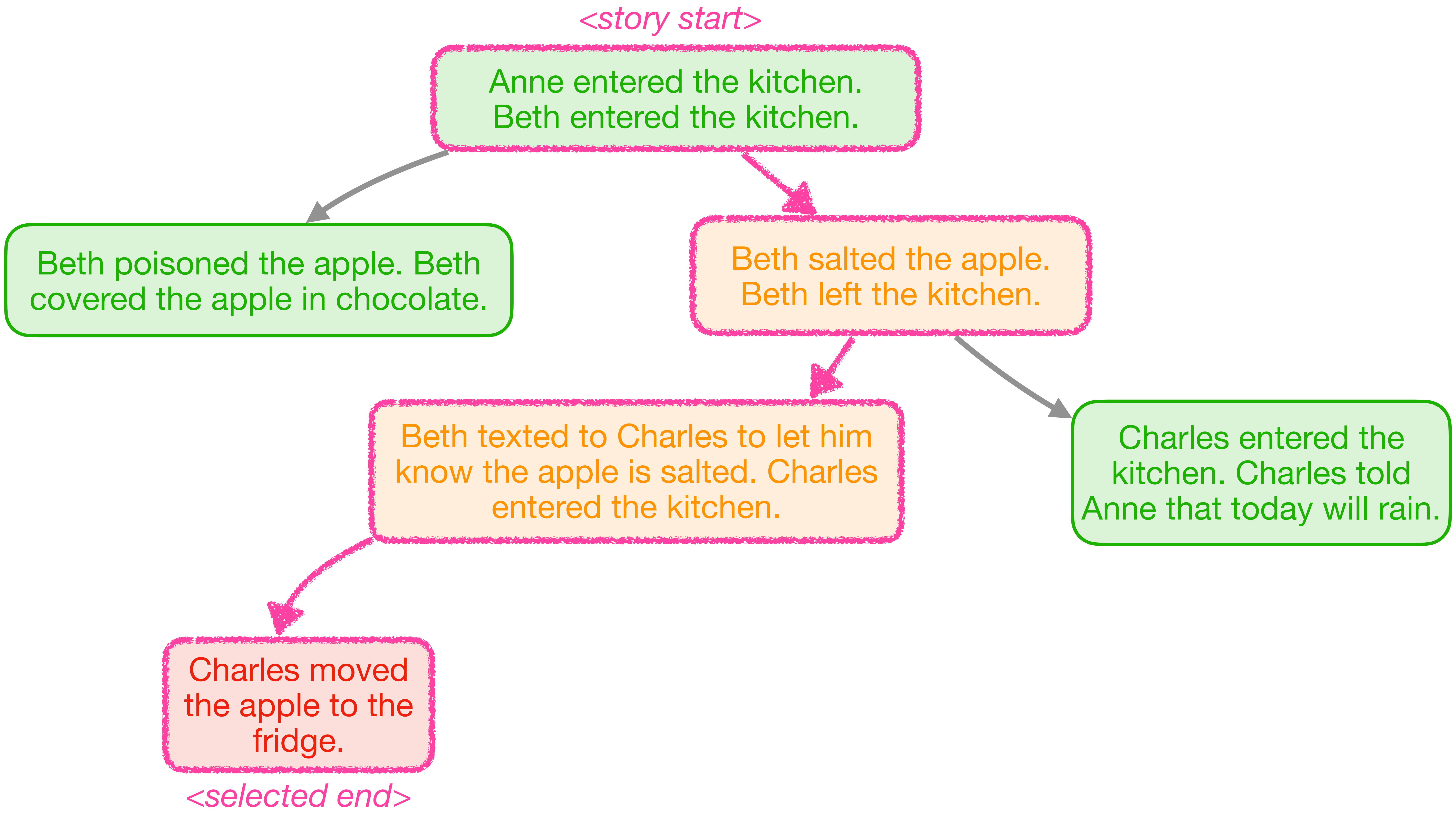
fridge ✓
table ✓
no ✗
yes ✗

acc = 0.5

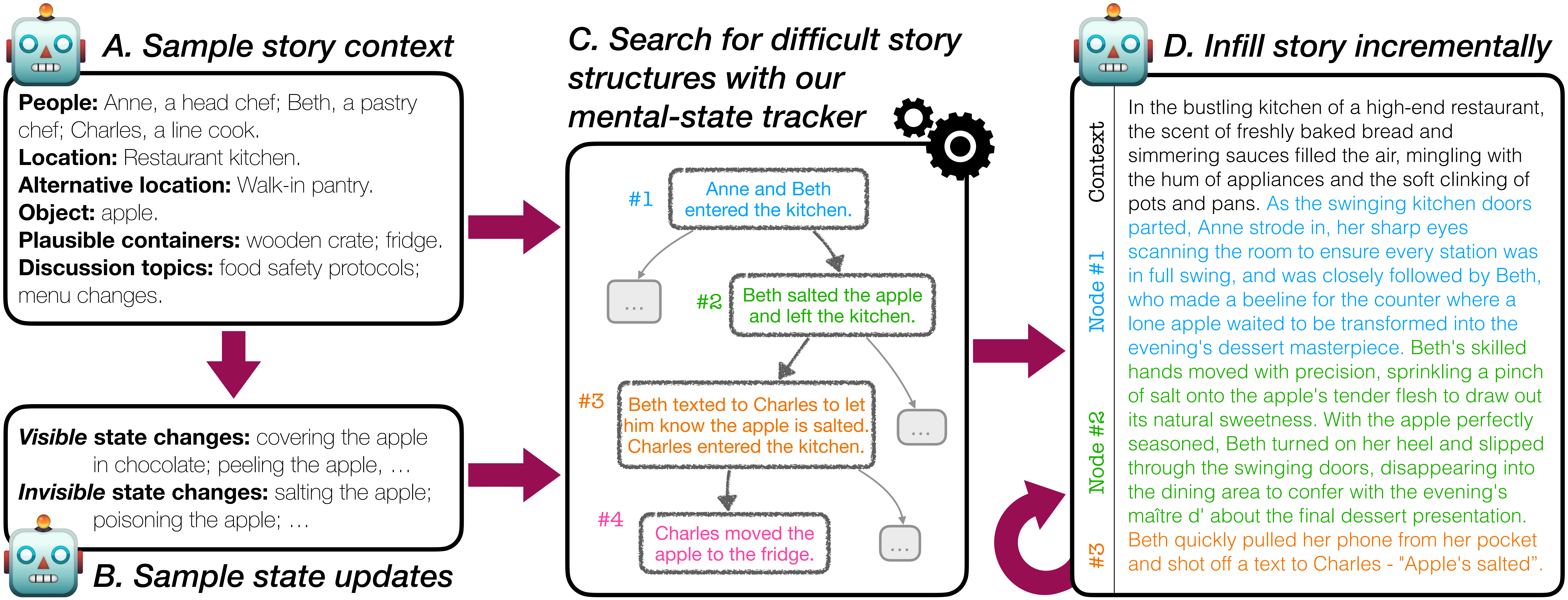
Adversarial story generation: *searching for difficult stories with A**



Adversarial story generation: *searching for difficult stories with A**



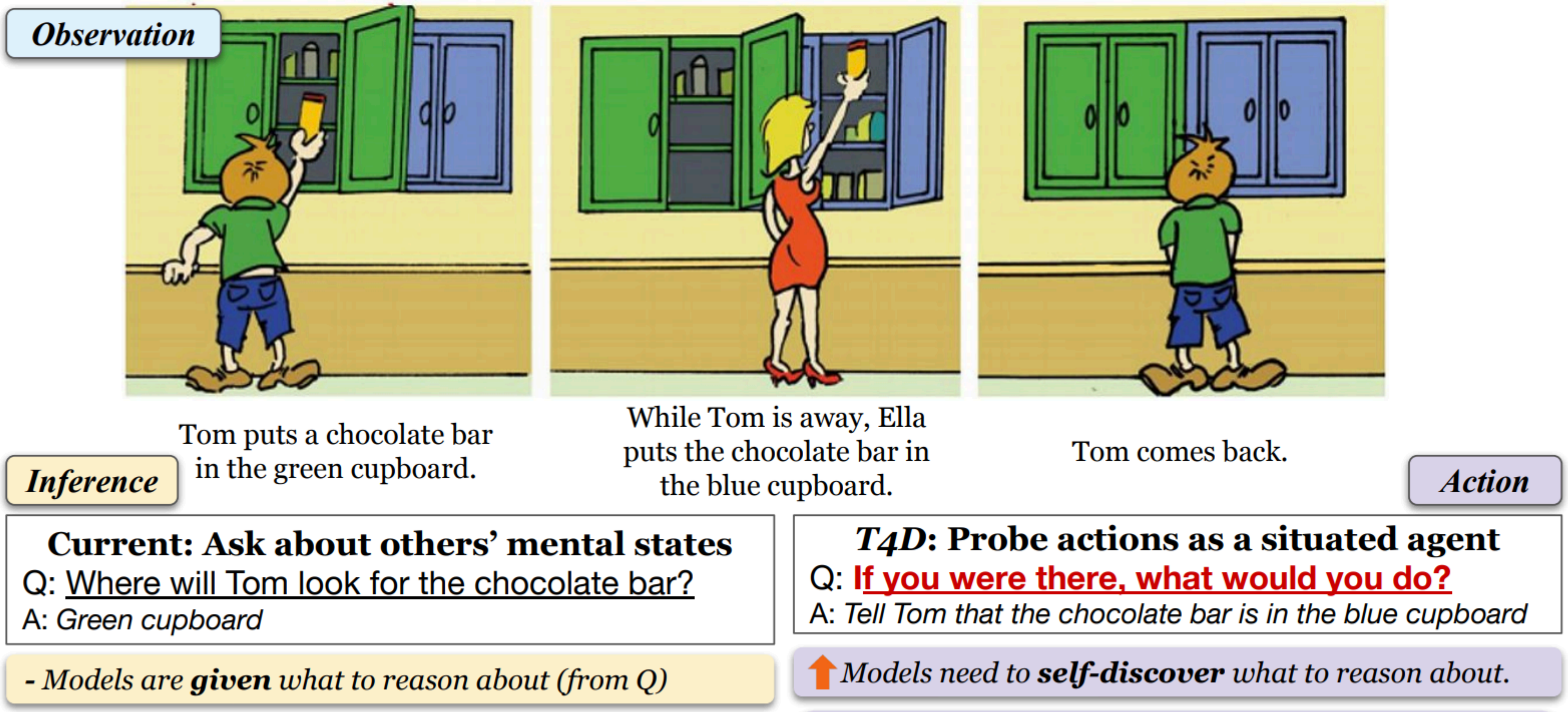
TrackTheMind: *full setup*



[A subset of] TrackTheMind results

TrackTheMind action set: $\{a_{\text{enter}}, a_{\text{leave}}, \ldots$	GPT-4o Accuracy
$\ldots, \boxed{a_{\text{moveObjContainer}}}\}$	.40
$\ldots, \boxed{a_{\text{updateObjState}}}\}$	.17
$\ldots, \boxed{a_{\text{moveObjContainer}}}, \boxed{a_{\text{updateObjState}}}\}$	.35
$\ldots, a_{\text{moveObjContainer}}, \boxed{a_{\text{moveObjRoom}}}\}$	.05
$\ldots, a_{\text{moveObjContainer}}, \boxed{a_{\text{info-*}}}\}$	.36
$\ldots, a_{\text{moveObjContainer}}, a_{\text{moveObjRoom}}, \boxed{a_{\text{info-*}}}\}$	.24
$\ldots, a_{\text{moveObjContainer}}, a_{\text{moveObjRoom}}, a_{\text{chitChat-*}}, \boxed{a_{\text{info-*}}}\}$	.71
$\ldots, \boxed{a_{\text{chitChat-private}}}\}$	.76
$\ldots, \boxed{a_{\text{chitChat-public}}}\}$	.46

Comment: LLMs are inconsistent, part 1000



Zhou, Pei, et al. HOW FAR ARE LARGE LANGUAGE MODELS FROM AGENTS WITH THEORY-OF-MIND?. 2024

Improving Reasoning

At Training time vs. Inference time

TrackTheMind: fine-tuning for improving reasoning

What if we used all the data we generated to teach a small model to be a better theory of mind reasoner?

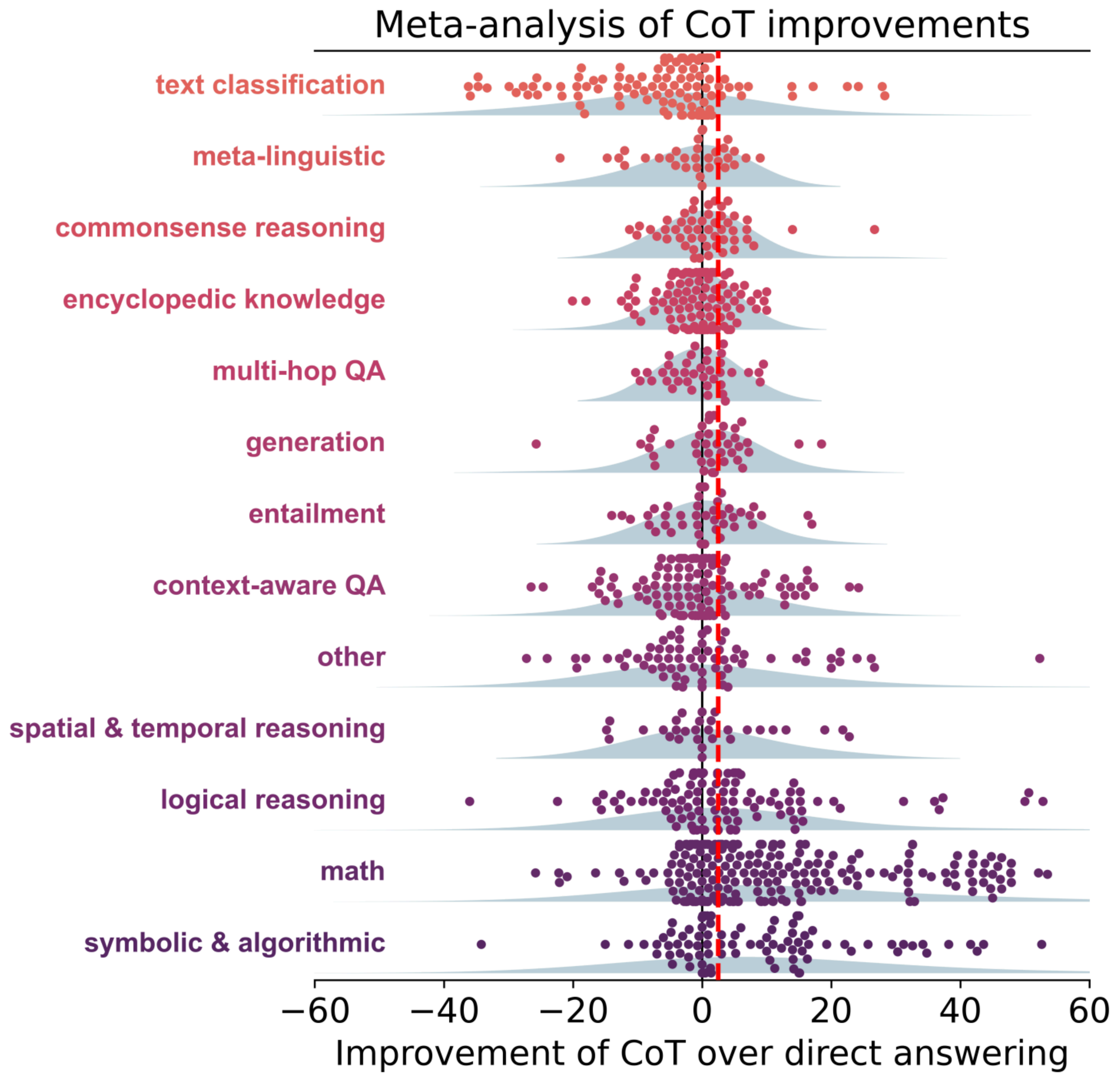
	ToMi	Hi-ToM	BigToM	OpenToM (F1)	FANToM
Llama-3.1 8B Instruct	68%	30%	75%	.39	0.3%
TRACKTHEMIND-8B	**94%** (+26)	**52%** (+22)	**79%** (+4)	**.42** (+.03)	0.5% (+0.02)

Example of a possible action plan when even frontier models cannot generate good data for knowledge distillation!
(See also Jung et al., 2023 in the previous lecture)

Inference-time algorithms for improving reasoning

- *Improving reasoning at training time*
  - **Pros:** you hopefully finish with an overall better model!
  - **Cons:** you need to find good data, which may be difficult; you might overfit
- *Improving reasoning at inference time*
  - **Pros:** does not require training data
  - **Cons:** possibly high cost we pay every time we want to run an algorithm; may not generalize too well

Inference-time algorithms for improving reasoning: CoT does not seem to be the holy grail



SPRAGUE ET AL 2024. TO COT OR NOT TO COT? CHAIN-OF-THOUGHT HELPS MAINLY ON MATH AND SYMBOLIC REASONING.

Inference-time algorithm example for improving theory of mind through **symbolic representations**

Integrating Belief Graphs to LLMs

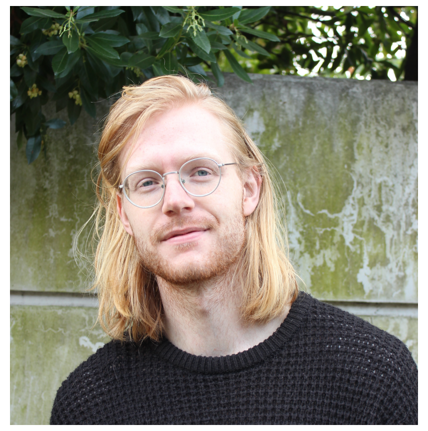
Minding Language Models' (Lack of) Theory of Mind:

A Plug-and-Play Multi-Character Belief Tracker

🏆 **Outstanding Paper Award at ACL 2023**

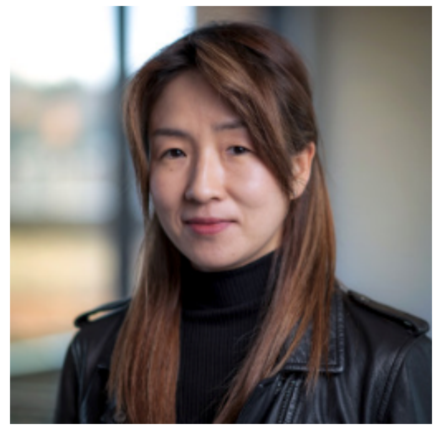
Melanie Sclar

Sachin Kumar



Peter West

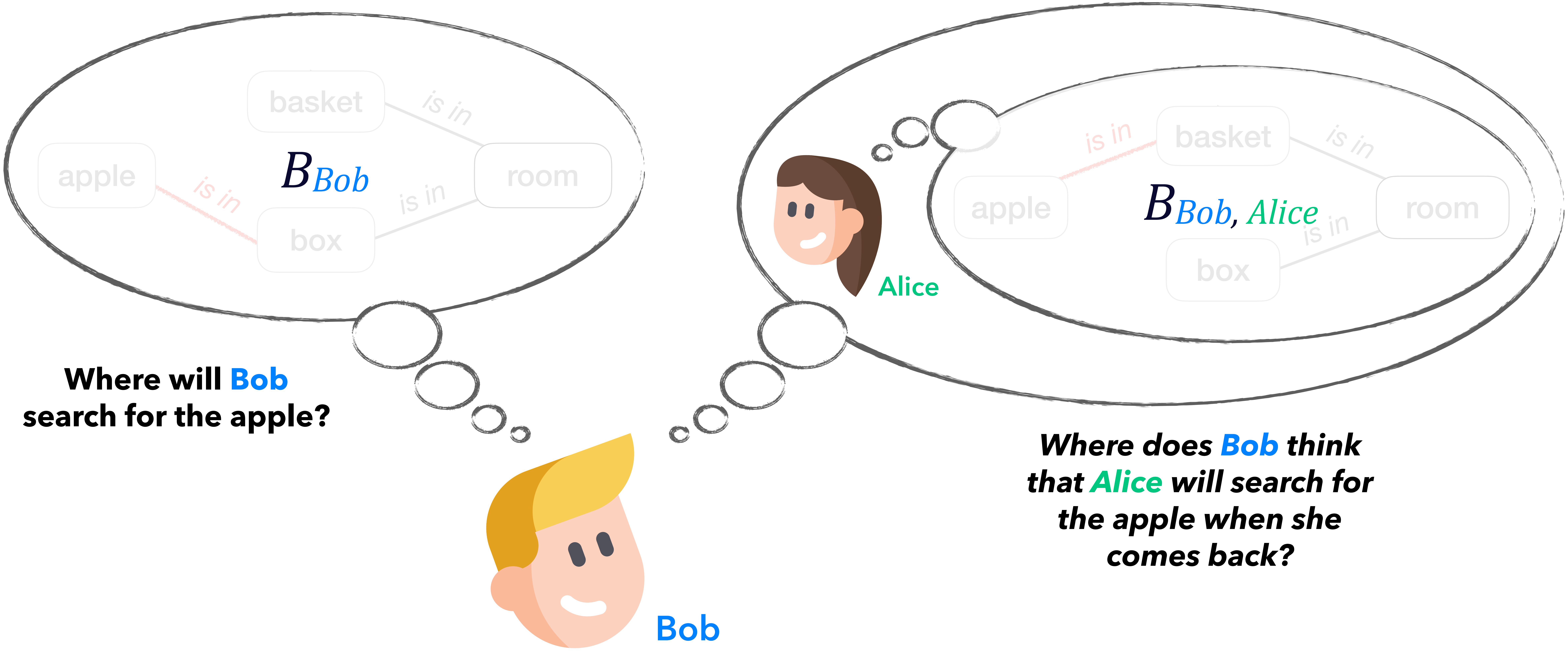
Alane Suhr



Yejin Choi

Yulia Tsvetkov

Graphical Representations of Local Context



Where will Bob search for the apple?

Where does Bob think that Alice will search for the apple when she comes back?

Symbolic ToM Overview

1. Detect entities in question, retrieve belief graph and perform recursion over the question

2. Retrieve sentences captured by the graph

3. Feed to Language Model

Results: Out-of-Domain Performance

Story Structure Generalization

	D_1	D_2	D_3
SymbolicToM + Off-the-shelf models			
Macaw-3B	89 (+81)	71 (+60)	70 (+41)
Flan-T5-XL	76 (-10)	96 (+46)	**100** (+33)
Flan-T5-XXL	93 (+24)	**100** (+41)	**100** (+49)
GPT3-Curie	84 (+48)	81 (+42)	73 (+16)
GPT3-Davinci	92 (+73)	91 (+66)	90 (+50)
GPT3.5	**100** (+99)	**100** (+99)	99 (+51)
GPT4	**100** (+42)	**100** (+38)	**100** (+4)
LLaMA-7B	99 (+82)	92 (+75)	88 (+71)
LLaMA-13B	78 (+52)	84 (+48)	84 (+47)
Supervised models			
TTT	49	65	78
Finetuned GPT3	51	68	32

ToM for this lecture

"I know that you believe you understand what you think I said,
but I'm not sure you realize that what you heard is not what I meant."

Alan Greenspan