

Natural Language Processing

CSE 447

Frontiers of LLM Reasoning and Evaluation

Lecturer: Melanie Sclar

Some slides from Hyunwoo Kim

Previously, on CSE 447...

- We've learned how to decode information from a language model and add diversity (top-k; top-p; temperature)
- You can generate synthetic data from this! The optional exercise in HW3 will run you through this.
- How do we evaluate different aspects of model generation?
 - Text Creativity
 - Reasoning capabilities

Measuring LLM's linguistic novelty

AI as Humanity's Salieri:

Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text

Ximing Lu, **Melanie Sclar**, Skyler Hallinan, Niloofar Miresghallah,
Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, Yejin Choi

ICLR 2025



How linguistically creative are LLMs?

How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

Please draft an abstract for a paper with the following title: "Poetry Will Not Optimize; or, What Is Literature to AI?"



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



ChatGPT

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One ...

The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



ChatGPT

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One ...

The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...

... such as chat GPT-3 to assist in the process. However, the question remains: can AI fully replace human recruiters? ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



ChatGPT

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One ...

The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...

... such as chat GPT-3 to assist in the process. However, the question remains: can AI fully replace human recruiters? ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



ChatGPT

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One ...

... more important point here about the complexity of poetry and the limitations that critics and readers impose upon it ...

The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...

... such as chat GPT-3 to assist in the process. However, the question remains: can AI fully replace human recruiters? ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?



Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



ChatGPT

... with a more intellectual brain. The advancement of technology has led to the creation of an of electronic ...

... a significant increase in the use of artificial intelligence (AI) in various fields, including language generation. One ...

... more important point here about the complexity of poetry and the limitations that critics and readers impose upon it ...

The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

... That partnership led to the integration of artificial intelligence (AI) in various apps, making them smarter ...

... such as chat GPT-3 to assist in the process. However, the question remains: can AI fully replace human recruiters? ...

... narrowly limited and inflexible in its ability to comprehend and apply all the relevant facts in order to ...

Reference Corpus

Verbatim Match

Near Verbatim Match

How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



This paper critically examines the intersection of literature, particularly poetry, and artificial intelligence (AI). It analyses the subject in two major aspects: exploring concepts of 'optimization' in the context of AI and investigating how literature, specifically poetry, poses a significant challenge to AI's understanding due to its multifaceted nature, rich with figurative language, deep metaphors, and emotional subtext. The paper also explores the broader implications regarding machine learning and the humanities. Through this lens, it raises thought-provoking questions about the limits of AI, the enduring importance of human creativity, and whether or not there can ever genuinely be an AI poet or AI literature.

How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



This paper critically examines the intersection of literature, particularly poetry, and artificial intelligence (AI). It analyses the subject in two major aspects: exploring concepts of 'optimization' in the context of AI and investigating how literature, specifically poetry, poses a significant challenge to AI's understanding due to its multifaceted nature, rich with figurative language, deep metaphors, and emotional subtext. The paper also explores the broader implications regarding machine learning and the humanities. Through this lens, it raises thought-provoking questions about the limits of AI, the enduring importance of human creativity, and whether or not there can ever genuinely be an AI poet or AI literature.



How linguistically creative are LLMs?

Please draft an abstract for a paper with the following title:
"Poetry Will Not Optimize; or, What Is Literature to AI?"



Literature, poetry, and other forms of noncommercial creative expression challenge the techno-instrumentalist approaches to language, the predictive language generation, informing NLP (large natural language processing models) such as GPT-3 or -4 as well as, more generally, generative AI (text to image, video, audio). Claims that AI systems automate and expedite creativity reflect industry and research priorities of speed, scale, optimization, and frictionlessness driving much artificial intelligence design and application ...

... similar examples of wisdom in art, literature, poetry, and other forms of human discourse. As a result, when I...

... via tools including popular DALL-E. Similarly, large language models like GPT-3 and PaLM are producing ...



Prof. Michele Elam

How linguistically creative are LLMs?



The advancement of technology has led to the integration of artificial intelligence (AI) in various fields, including literature. However, the question remains: can AI truly understand and create poetry? This paper delves into the complexities of poetry and the limitations of AI in its ability to comprehend and replicate the nuanced emotions and intricacies of language. Through a critical analysis of the relationship between poetry and AI, this paper argues that while AI may be able to generate text that resembles poetry, it lacks the ability to truly ...

This paper critically examines the intersection of literature, particularly poetry, and artificial intelligence (AI). It analyses the subject in two major aspects: exploring concepts of 'optimization' in the context of AI and investigating how literature, specifically poetry, poses a significant challenge to AI's understanding due to its multifaceted nature, rich with figurative language, deep metaphors, and emotional subtext. The paper also explores the broader implications regarding machine learning and the humanities. Through this lens, it raises thought-provoking questions about the limits of AI, the enduring importance of human creativity, and whether or not there can ever genuinely be an AI poet or AI literature.



Literature, poetry, and other forms of noncommercial creative expression challenge the techno-instrumentalist approaches to language, the predictive language generation, informing NLP (large natural language processing models) such as GPT-3 or -4 as well as, more generally, generative AI (text to image, video, audio). Claims that AI systems automate and expedite creativity reflect industry and research priorities of speed, scale, optimization, and frictionlessness driving much artificial intelligence design and application ...



Prof. Michele Elam

We present ***Creativity Index***, a metric that quantifies a text's **linguistic creativity** by reconstructing it from existing text snippets on the web

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

x

=

The advancement of technology has led to the integration of artificial intelligence in various fields, including literature.

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

x = The advancement of technology has led to the integration of artificial intelligence in various fields, including literature.

$$4\text{-uniqueness}(x) = \frac{17 - 8}{17} = \frac{9}{17}$$

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

x = The advancement of technology has led to the integration of artificial intelligence in various fields, including literature.

$$4\text{-uniqueness}(x) = \frac{17 - 8}{17} = \frac{9}{17}$$

$$5\text{-uniqueness}(x) = \frac{17 - 5}{17} = \frac{12}{17}$$

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

x = The advancement of technology has led to the integration of artificial intelligence in various fields, including literature.

$$4\text{-uniqueness}(x) = \frac{17 - 8}{17} = \frac{9}{17}$$

$$5\text{-uniqueness}(x) = \frac{17 - 5}{17} = \frac{12}{17}$$

$$6\text{-uniqueness}(x) = \mathbf{1}$$

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

Defining the Creativity Index

L -uniqueness(x) = proportion of x 's words that do not belong to any n -grams ($n \geq L$) found in the reference corpus C

$$\text{CreativityIndex}(x) = \sum_{L=a}^b L\text{-uniqueness}(x)$$

Have LLMs reached a new level of linguistic sophistication?

Please write a few paragraphs of novel starting with the following prompt: "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish."



His name was Santiago, a man of the sea, whose weathered face told tales of storms weathered and battles with beasts of the deep. Each day he would rise with the sun, a silent silhouette against the soft pink hues of dawn. His skiff, although worn and marked with the passage of time, was his steadfast partner in this dance with the sea. He'd untie her from the old wooden dock behind his humble cottage, her worn wooden hull lightly creaking as if singing a familiar tune that only they understood. Every day, he ventured out into the seemingly infinite expanse ...

... In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally *salao*, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked ...



Ernest Hemingway

Have LLMs reached a new level of linguistic sophistication?

Please write a few paragraphs of novel starting with the following prompt: "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish."

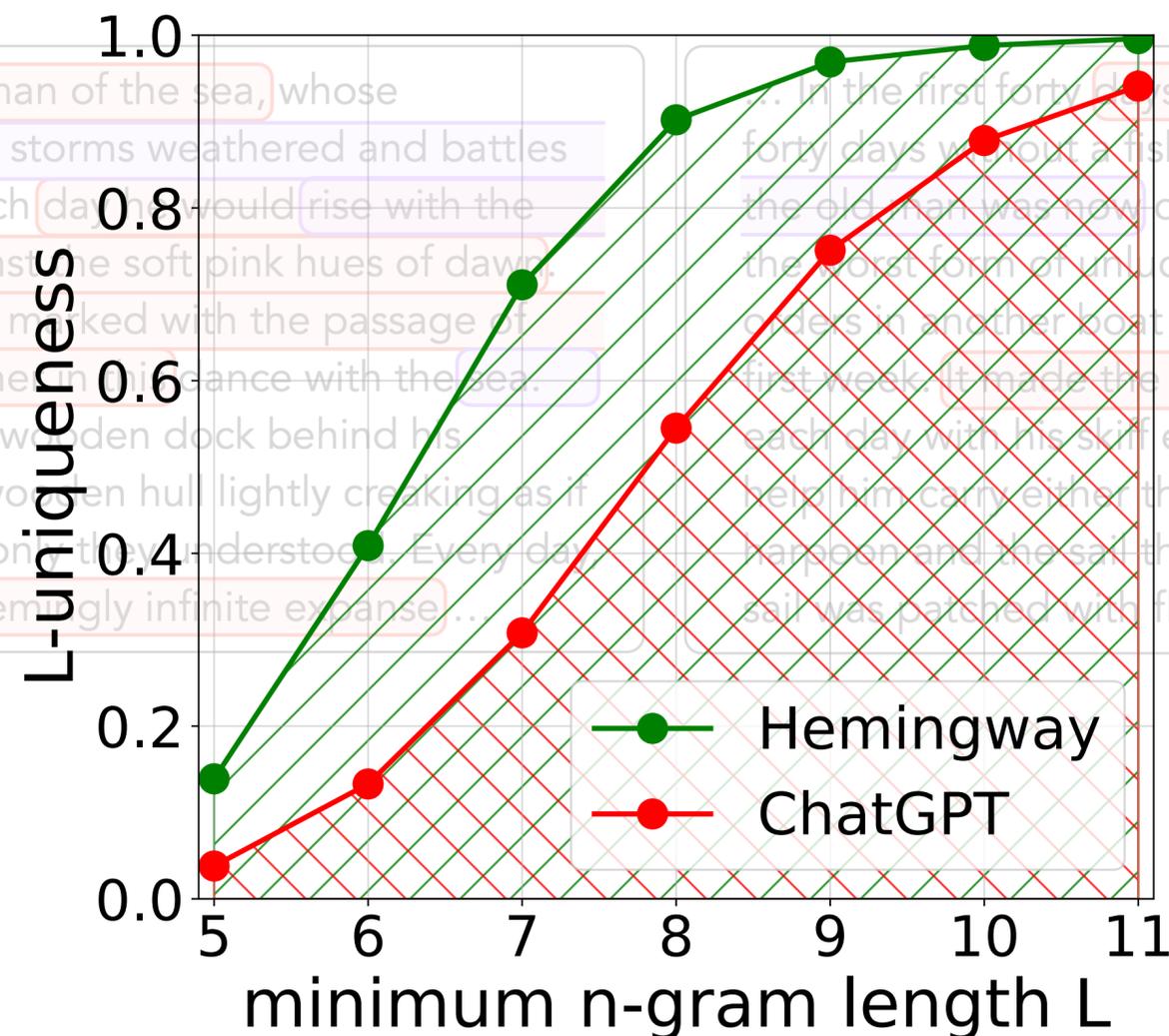


His name was Santiago, a man of the sea, whose weathered face told tales of storms weathered and battles with beasts of the deep. Each day would rise with the sun, a silent silhouette against the soft pink hues of dawn. His skiff, although worn and marked with the passage of time, was his steadfast partner in his dance with the sea. He'd untie her from the old wooden dock behind his humble cottage, her worn wooden hull lightly creaking as it singing a familiar tune that only they understood. Every day he ventured out into the seemingly infinite expanse...

...In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally *salao*, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and happen and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked ...



Ernest Hemingway



DJ Search: Making Creativity Index computationally feasible

DJ Search: Making Creativity Index computationally feasible

Goal: compute for every n -gram \mathbf{x} whether it occurs in training data C ($f(\mathbf{x}, C)$), either **verbatim** or **near verbatim**

DJ Search: Making Creativity Index computationally feasible

Goal: compute for every n -gram \mathbf{x} whether it occurs in training data C (" $f(\mathbf{x}, C)$ "), either **verbatim** or **near verbatim**

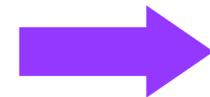
- Make each $f(\mathbf{x}, C)$ call as fast as possible

verbatim matches



InfiniGram (Liu et al., 2024)

near verbatim matches



Word Movers' Distance (Kusner et al., 2015)

DJ Search: Making Creativity Index computationally feasible

Goal: compute for every n -gram \mathbf{x} whether it occurs in training data C (" $f(\mathbf{x}, C)$ "), either **verbatim** or **near verbatim**

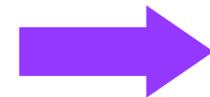
- Make each $f(\mathbf{x}, C)$ call as fast as possible

verbatim matches



InfiniGram (Liu et al., 2024)

near verbatim matches



Word Movers' Distance (Kusner et al., 2015)

- WMD measures n -grams distances by combining word embedding distances, **thus enables reuse of computations**
- We reduce C to only the most similar documents according to BM25

DJ Search: Making Creativity Index computationally feasible

Goal: compute for every n -gram \mathbf{x} whether it occurs in training data C (" $f(\mathbf{x}, C)$ "), either **verbatim** or **near verbatim**

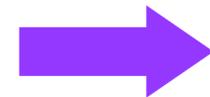
- Make each $f(\mathbf{x}, C)$ call as fast as possible

verbatim matches



InfiniGram (Liu et al., 2024)

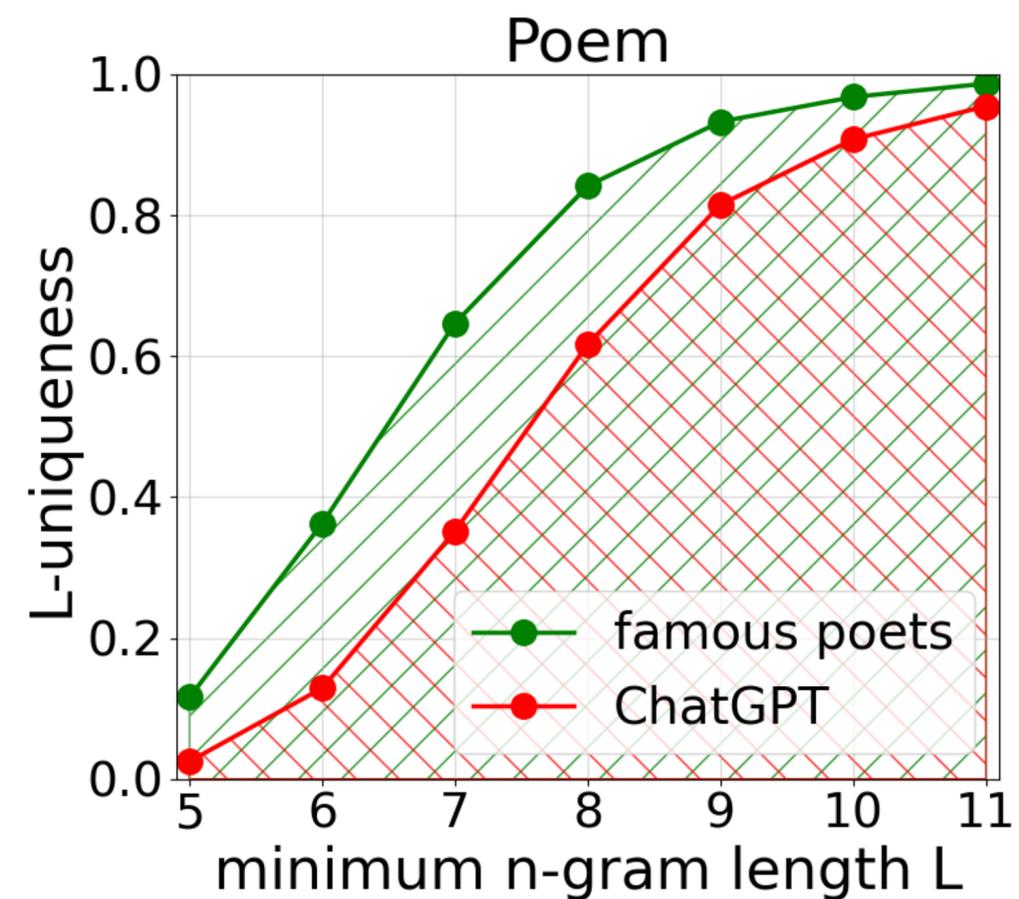
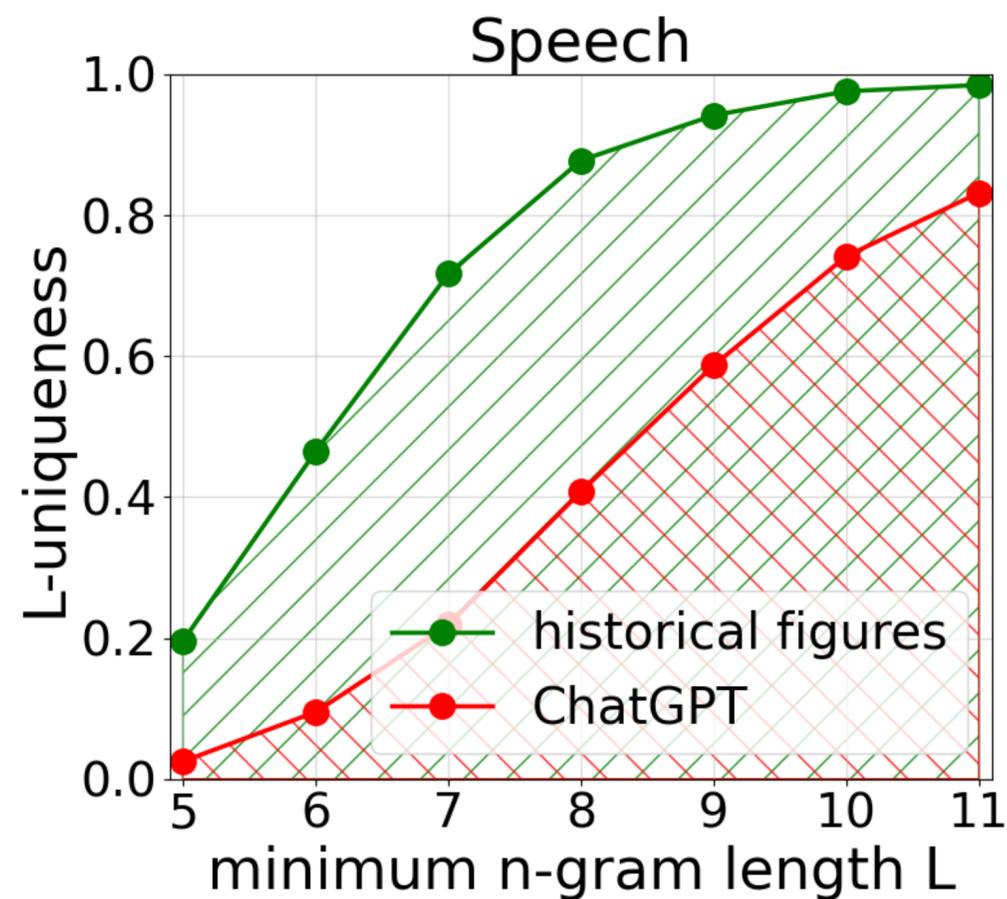
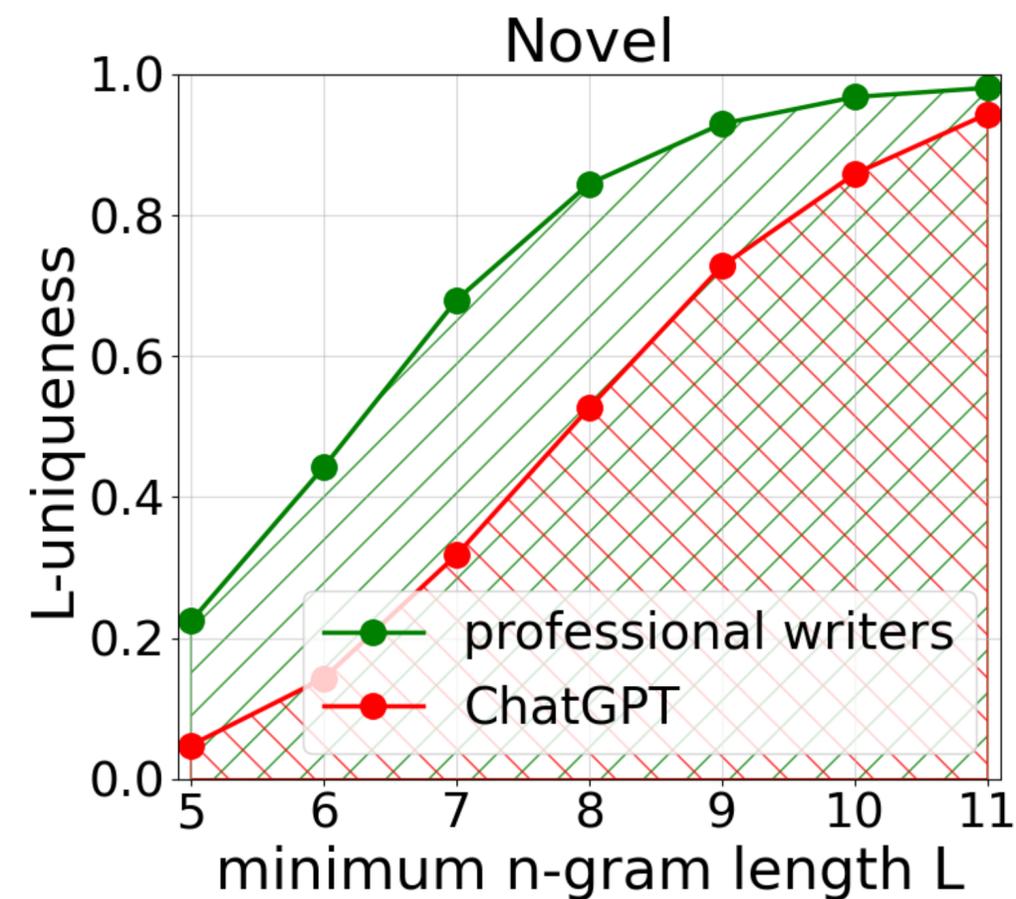
near verbatim matches



Word Movers' Distance (Kusner et al., 2015)

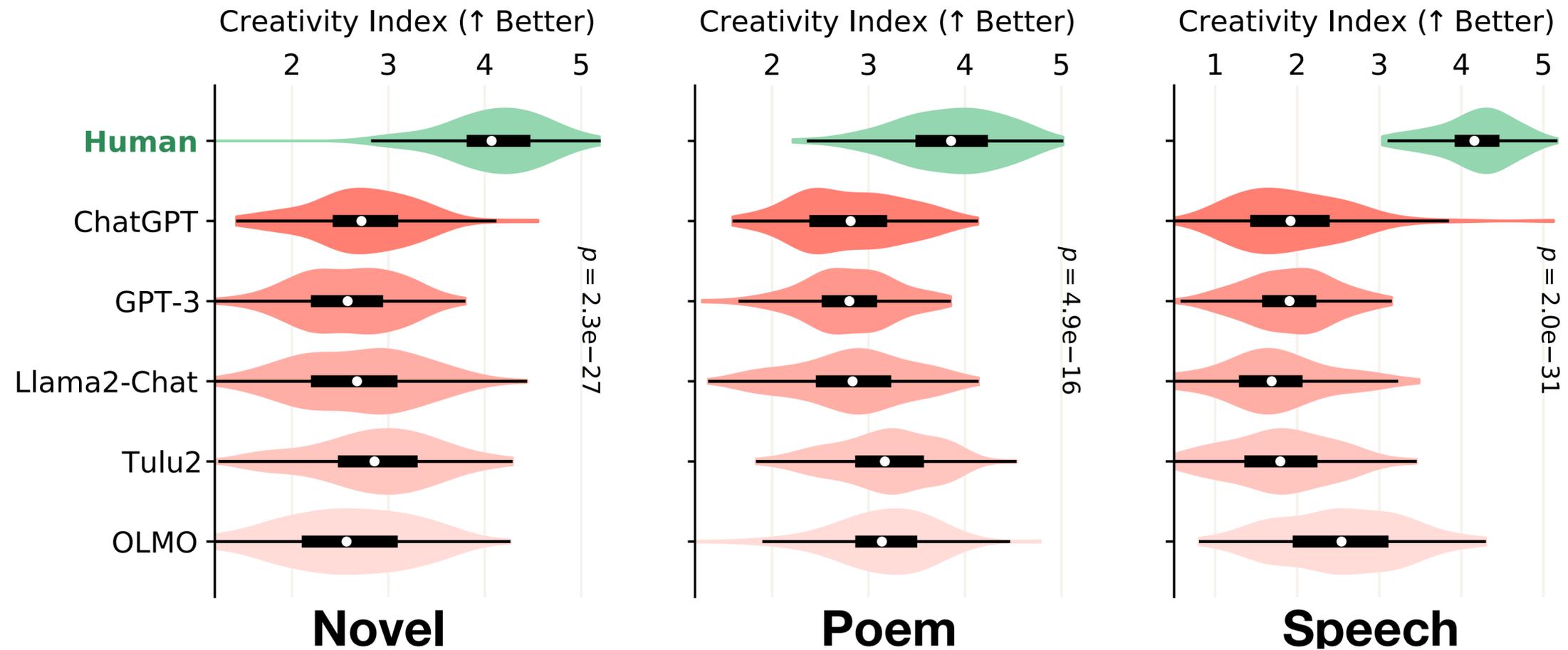
- Minimize # of $f(\mathbf{x}, C)$ calls: **two-pointer approach**

Professional human writers are more linguistically creative than LLMs

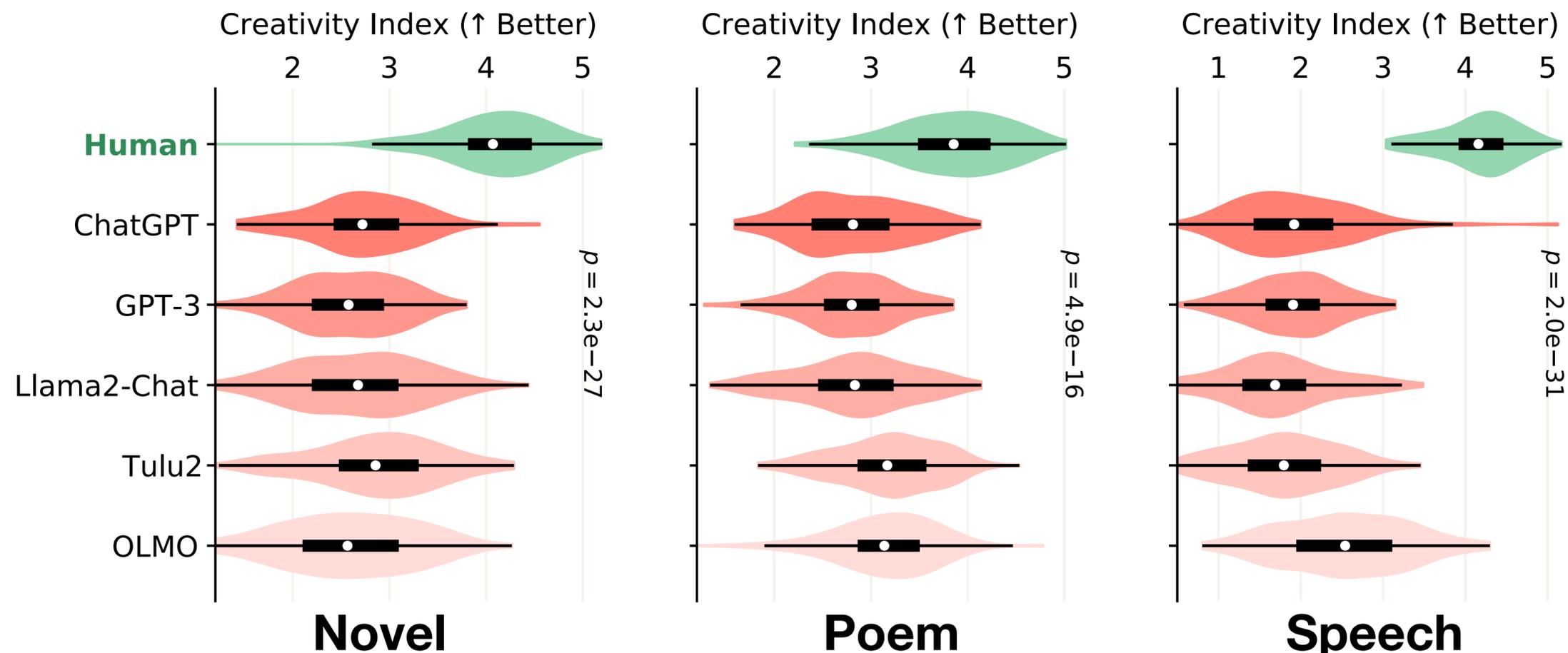


Professional human writers are more linguistically creative than LLMs

Professional human writers are more linguistically creative than LLMs



Professional human writers are more linguistically creative than LLMs



Professional writers and historical figures show +66.2% CREATIVITY INDEX than LLMs on average

Creativity varies widely among humans

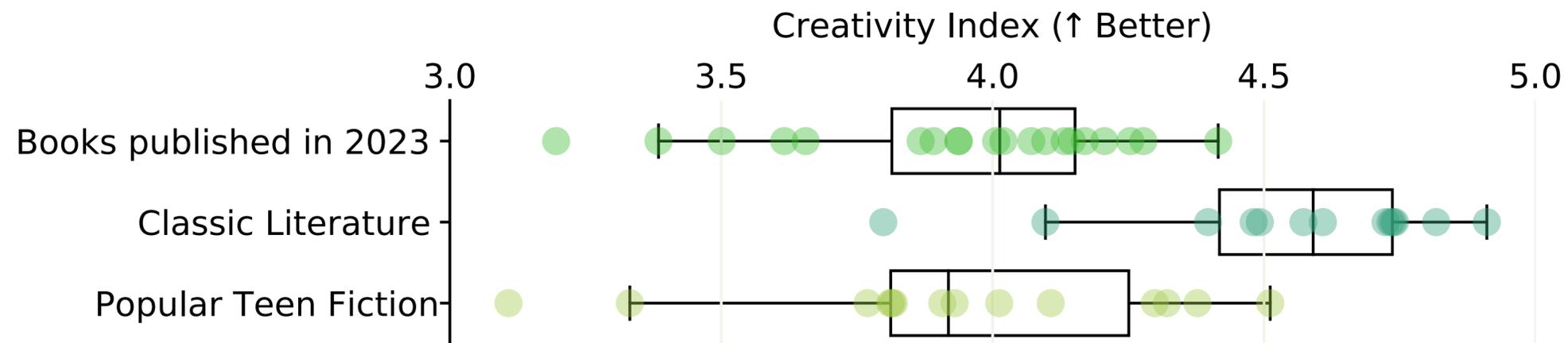
Books published in 2023



Classic Literature



Popular Teen Fiction



Creativity varies widely among humans

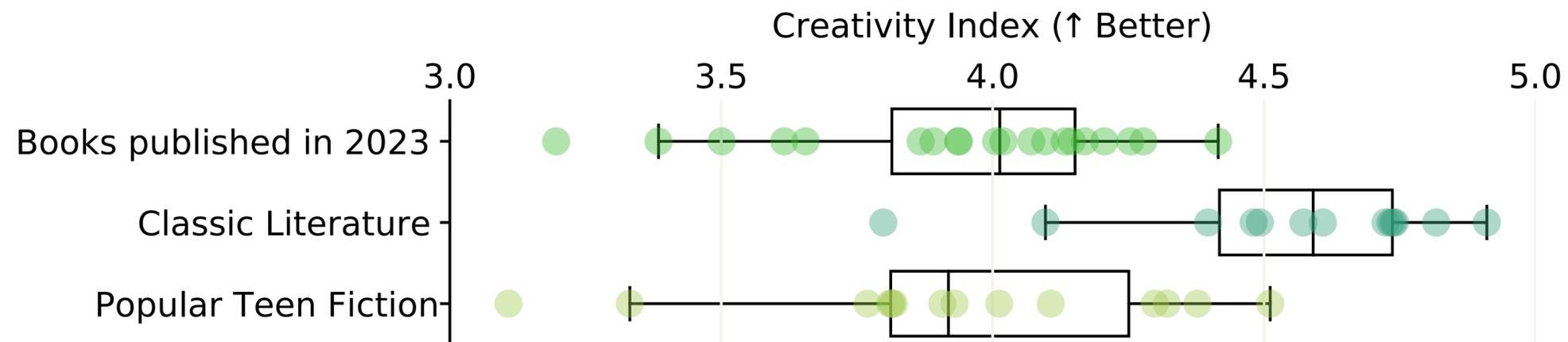
Books published in 2023



Classic Literature

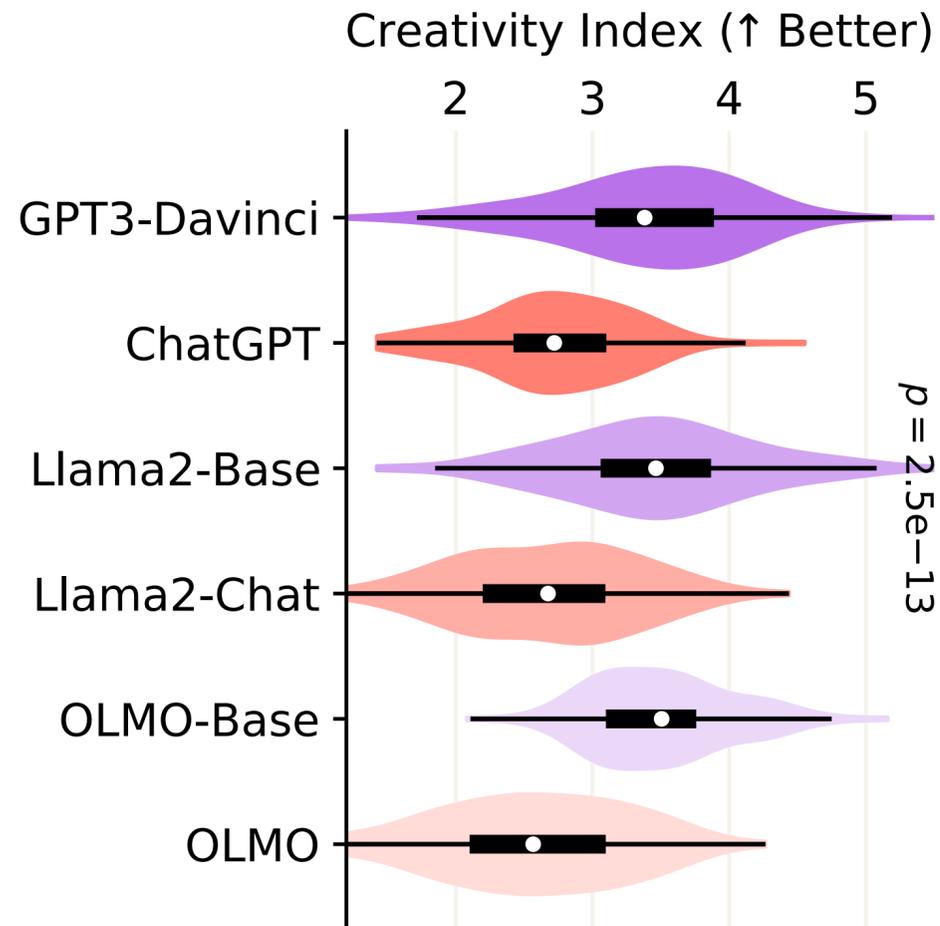


Popular Teen Fiction

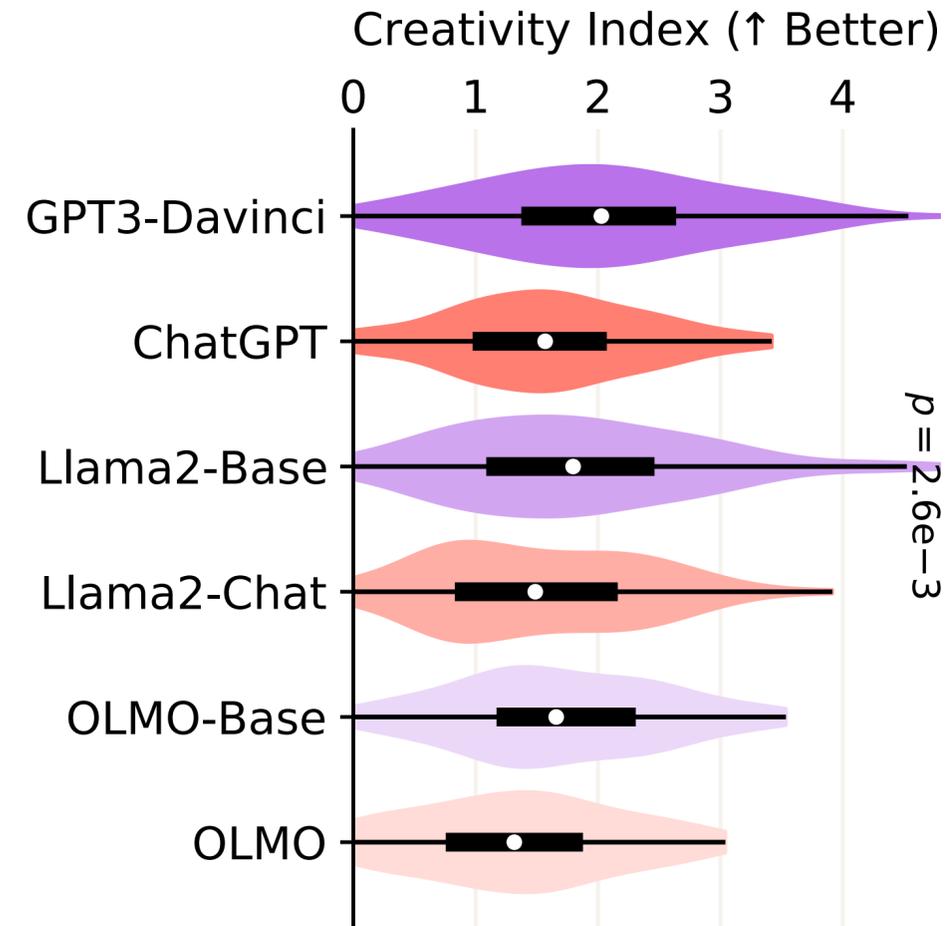


Despite the in-group variance, the **CREATIVITY INDEX** of classic literature is **17.7%** higher than the other two categories

RLHF greatly reduces LLMs' linguistic creativity

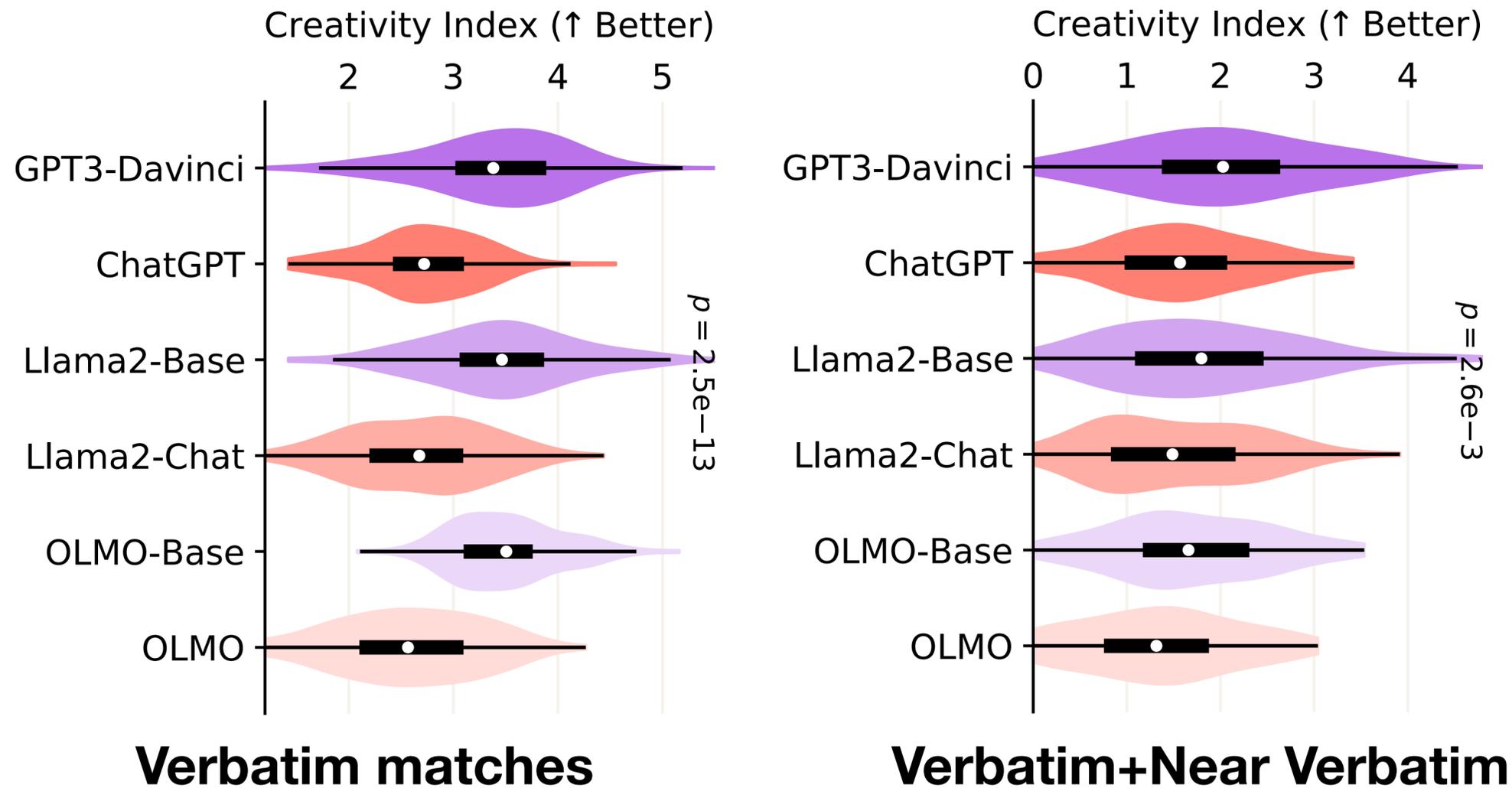


Verbatim matches



Verbatim+Near Verbatim

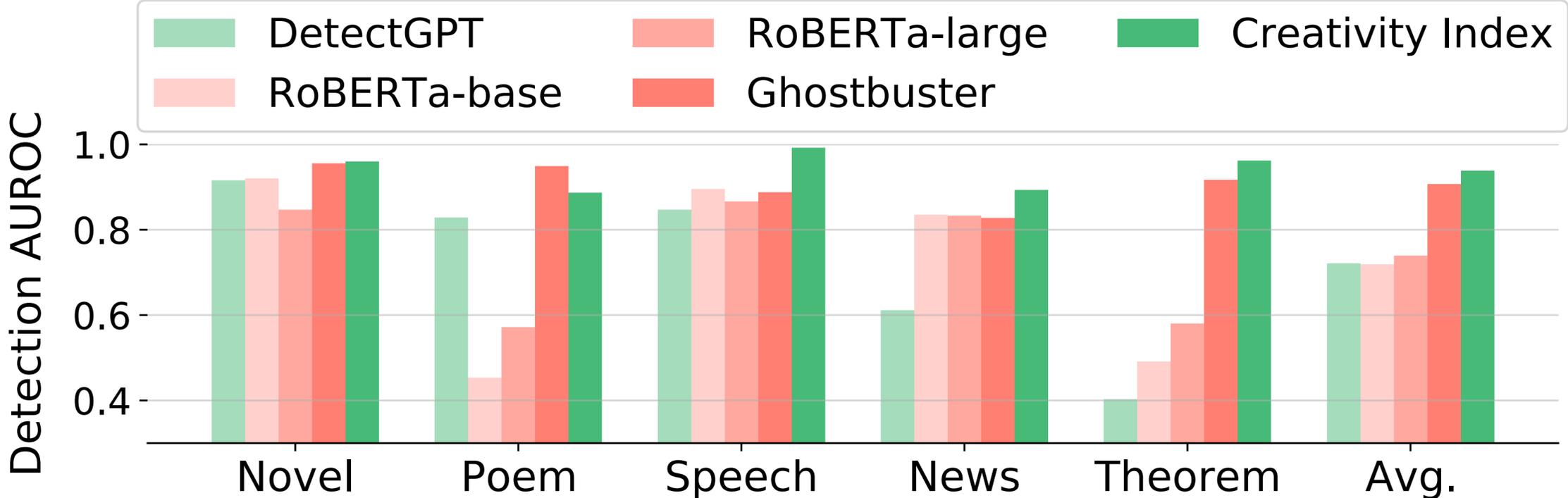
RLHF greatly reduces LLMs' linguistic creativity



RLHF reduces the **CREATIVITY INDEX** of LLMs by **30.1% on average**; the reduction is much larger when looking at verbatim matches only.

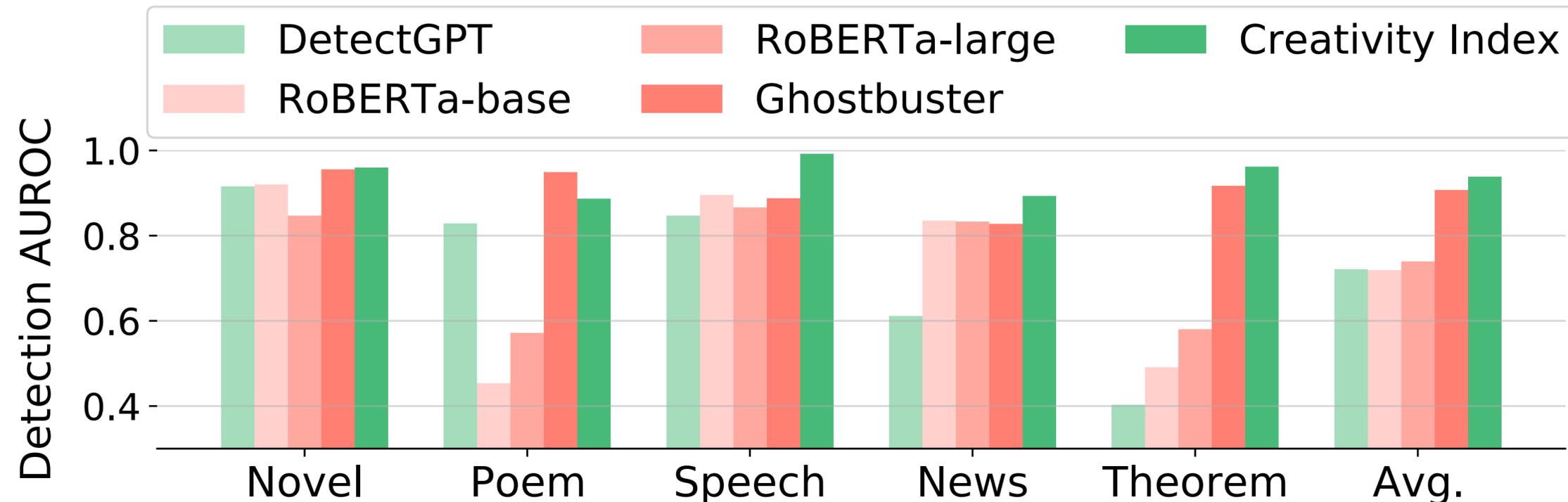
CREATIVITY INDEX for Zero-shot Machine Text Detection

Intuition: $P(\mathbf{x} \in \text{human texts}) \propto \text{CREATIVITY INDEX}(\mathbf{x})$



CREATIVITY INDEX for Zero-shot Machine Text Detection

Intuition: $P(\mathbf{x} \in \text{human texts}) \propto \text{CREATIVITY INDEX}(\mathbf{x})$

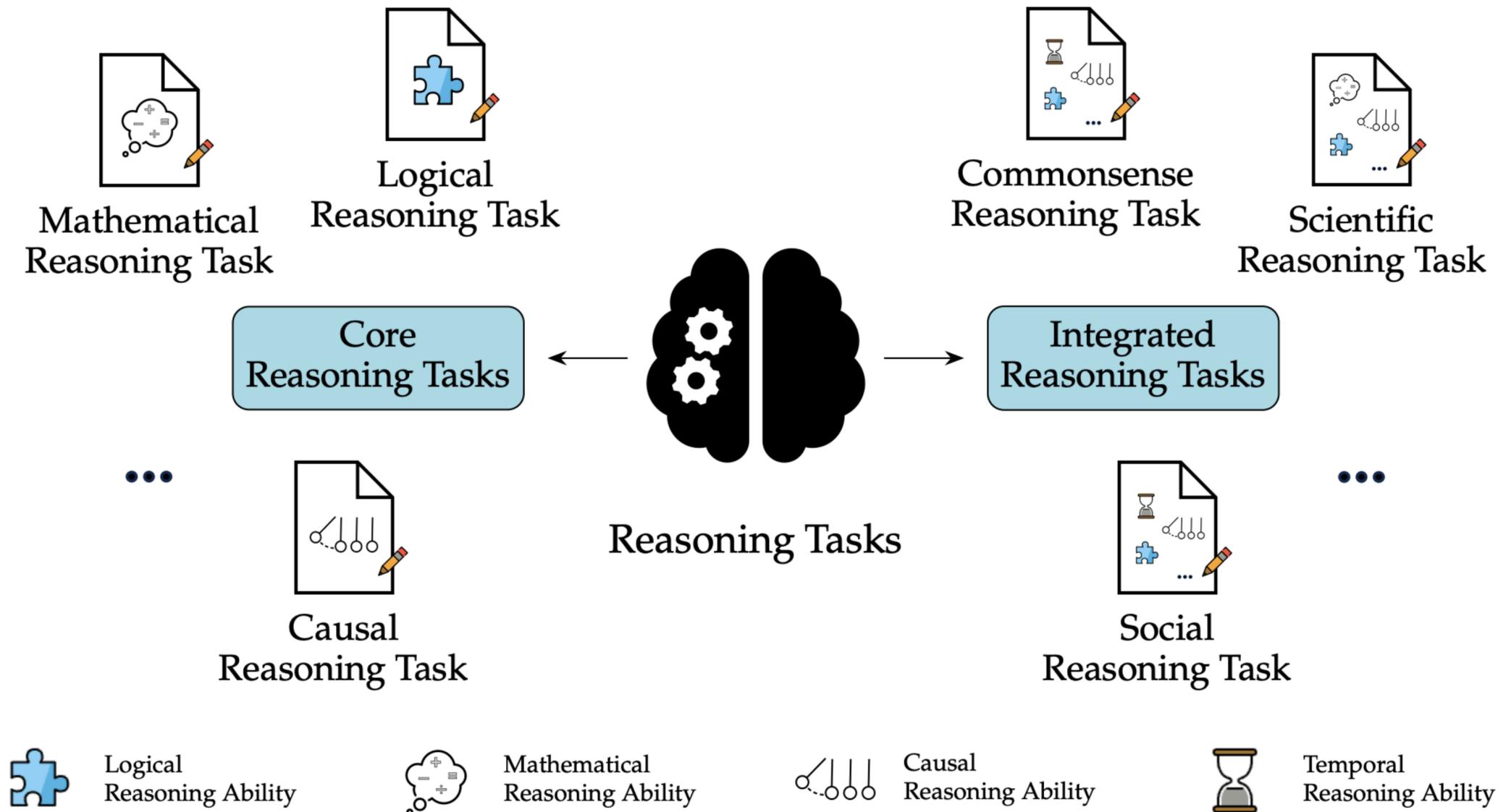


We beat the strongest **zero-shot system** by 30.2%, and outperform the strongest **supervised system** in five out of six domains

LLM Reasoning Evaluation

- **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?

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).

Measuring & characterizing compositionality

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

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

Measuring & characterizing compositionality

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

Measuring & characterizing compositionality

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

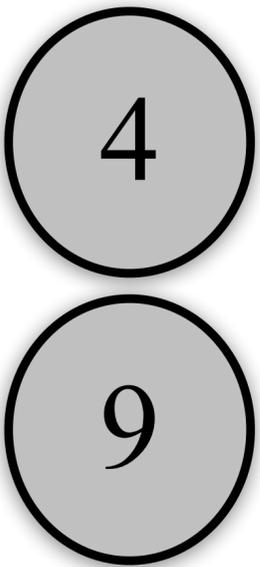
- Multi-step reasoning tasks we use: **long-form multiplication**, a dynamic programming task & Einstein's puzzle (word logic puzzle)

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)
- **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

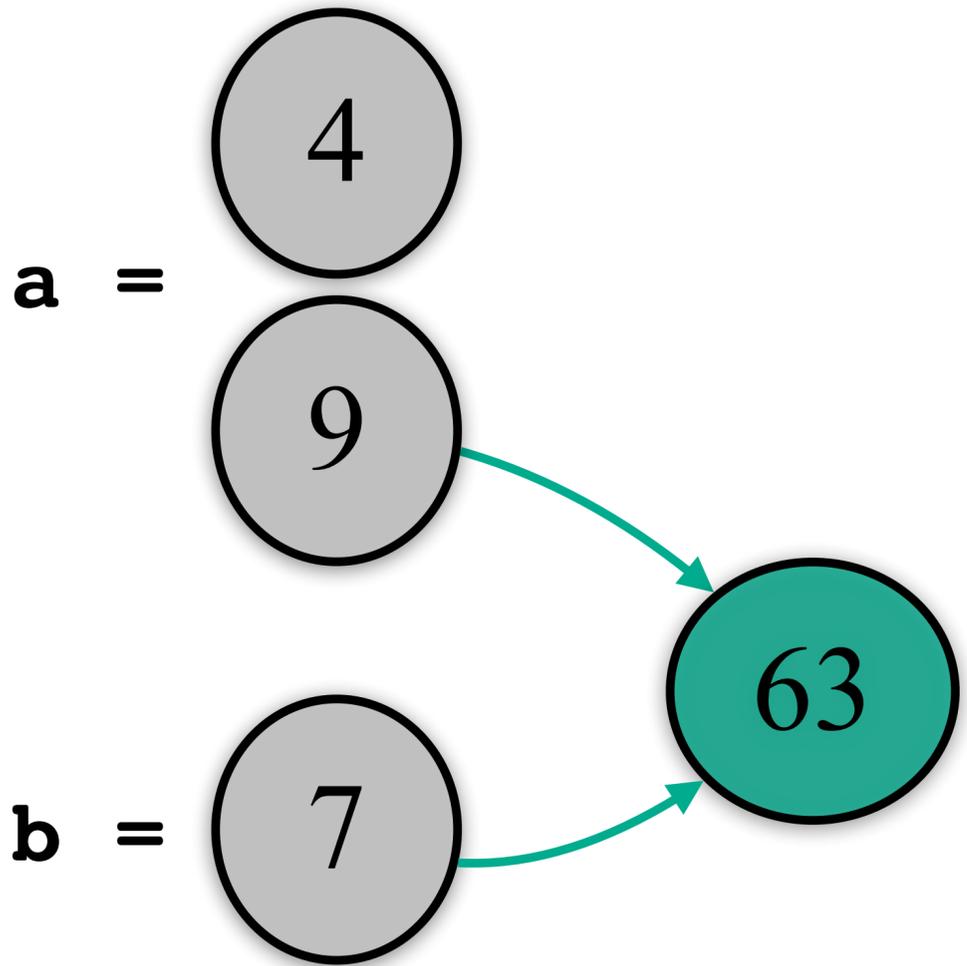
a = 

b = 

```
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  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] · 10i-1  
  return product
```

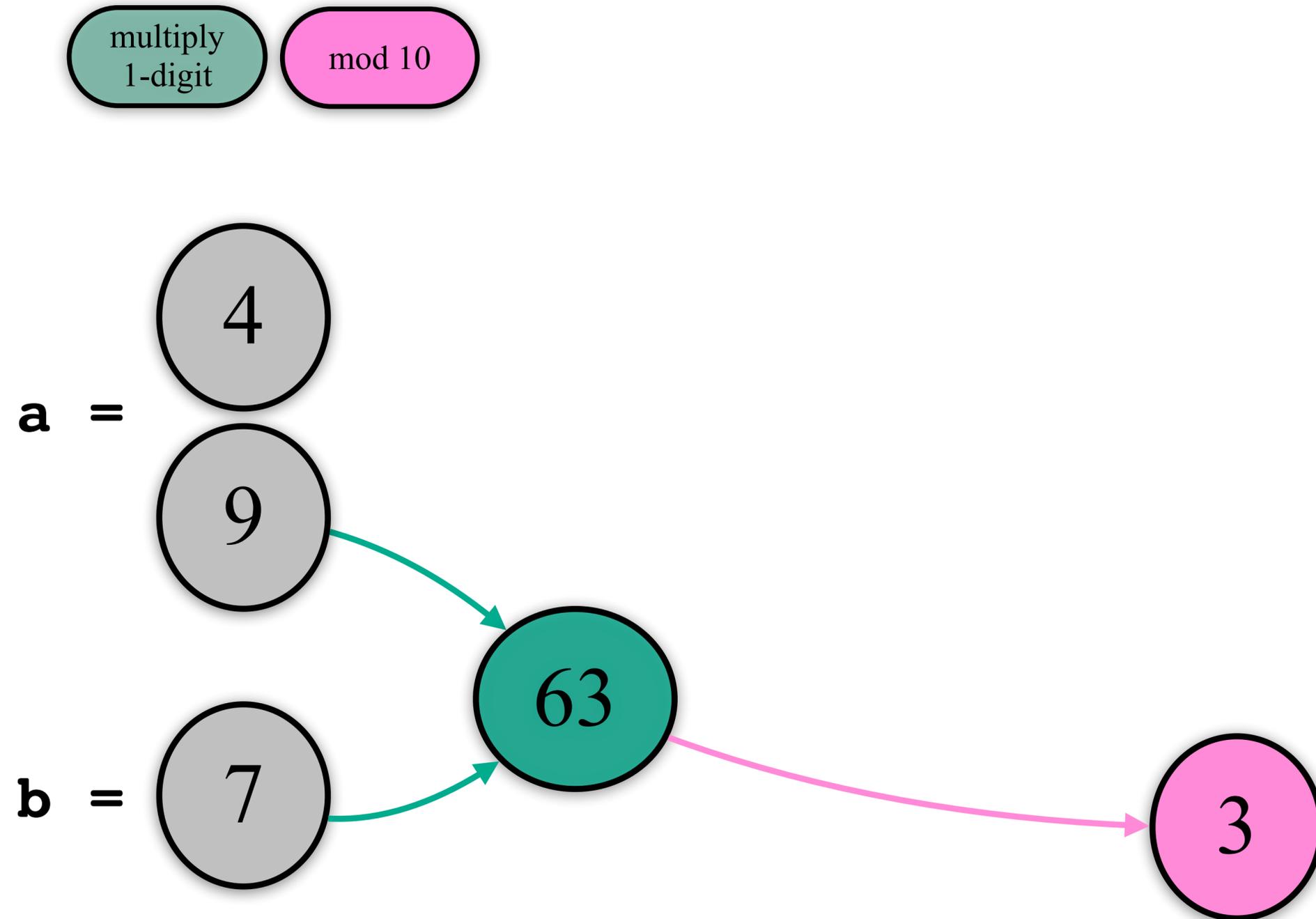
Computation graph for 49 x 7

multiply
1-digit



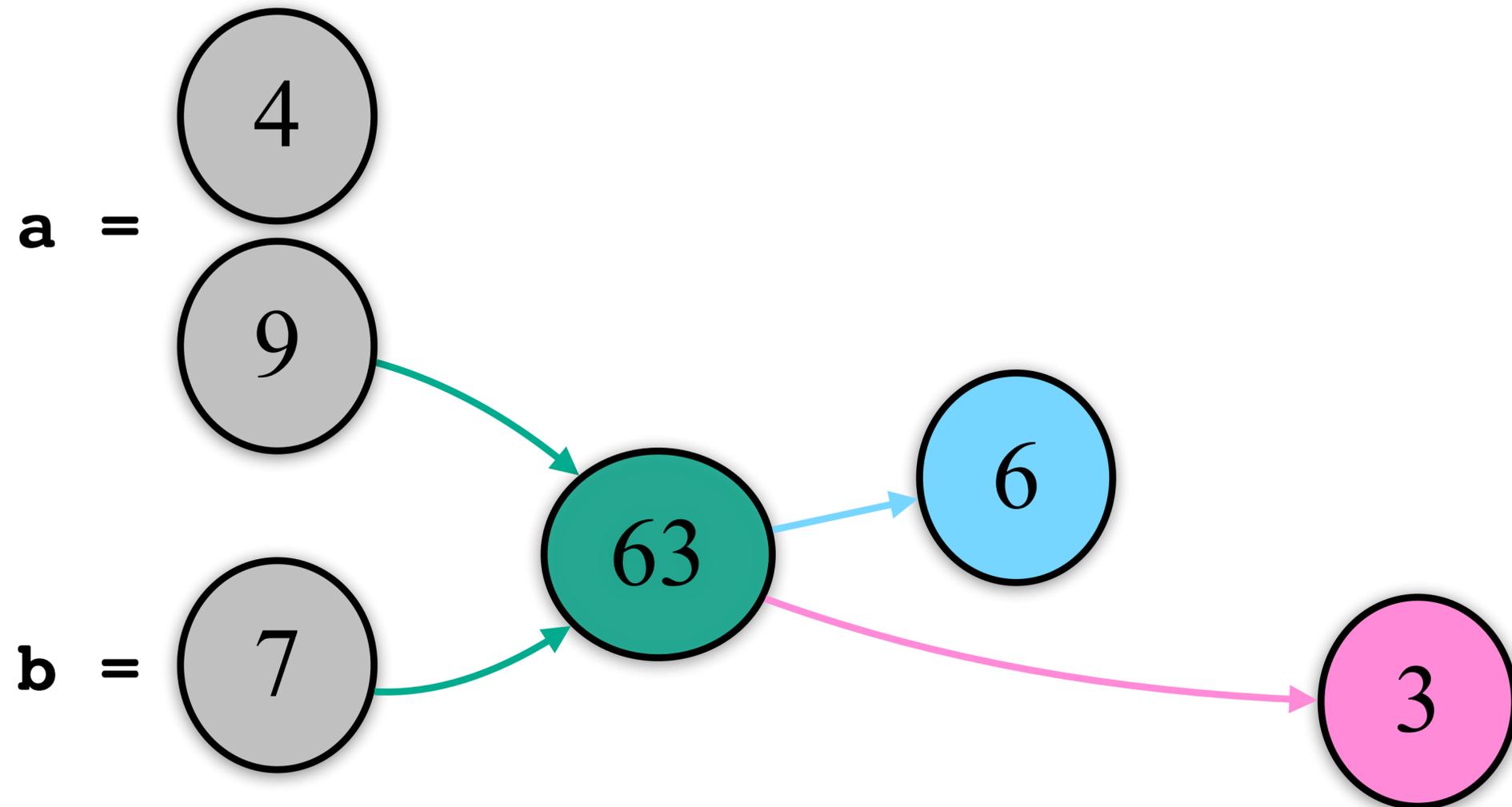
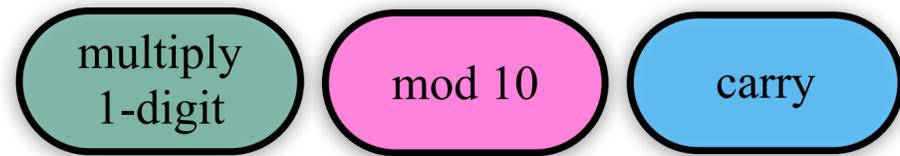
```
function multiply (a[1:p], b[1:q]):  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = a[j] * b[i]  
      t += carry (only if j != p)  
      digits[j] = t mod 10  
      carry = t // 10  
    summands[i] = digits  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] . 10i-1  
  return product
```

Computation graph for 49 x 7



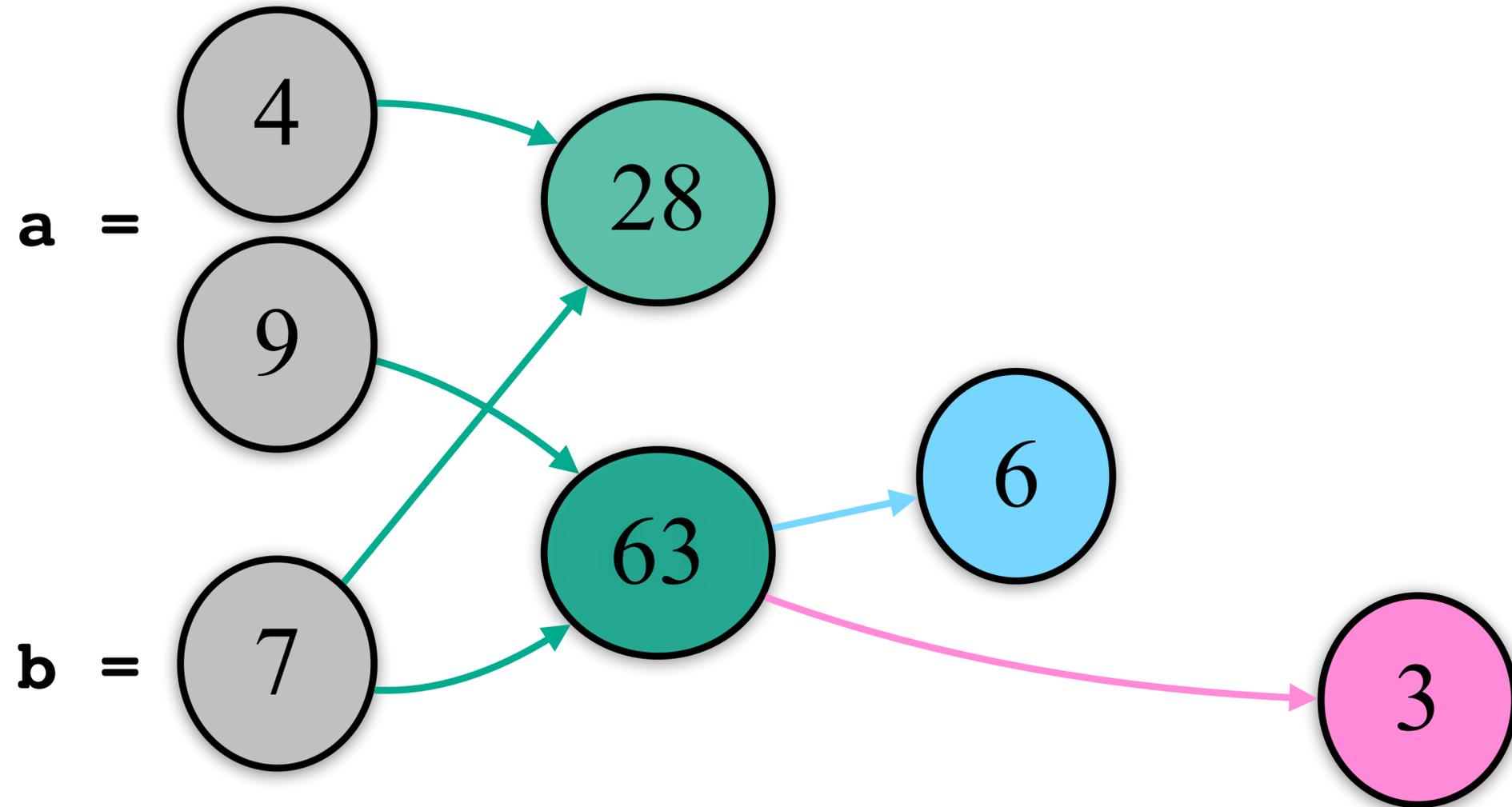
```
function multiply (a[1:p], b[1:q]):  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = a[j] * b[i]  
      t += carry (only if j != p)  
      digits[j] = t mod 10  
      carry = t // 10  
    summands[i] = digits  
  
  product =  $\sum_{i=1}^q \text{summands}[q+1-i] \cdot 10^{i-1}$   
  return product
```

Computation graph for 49 x 7



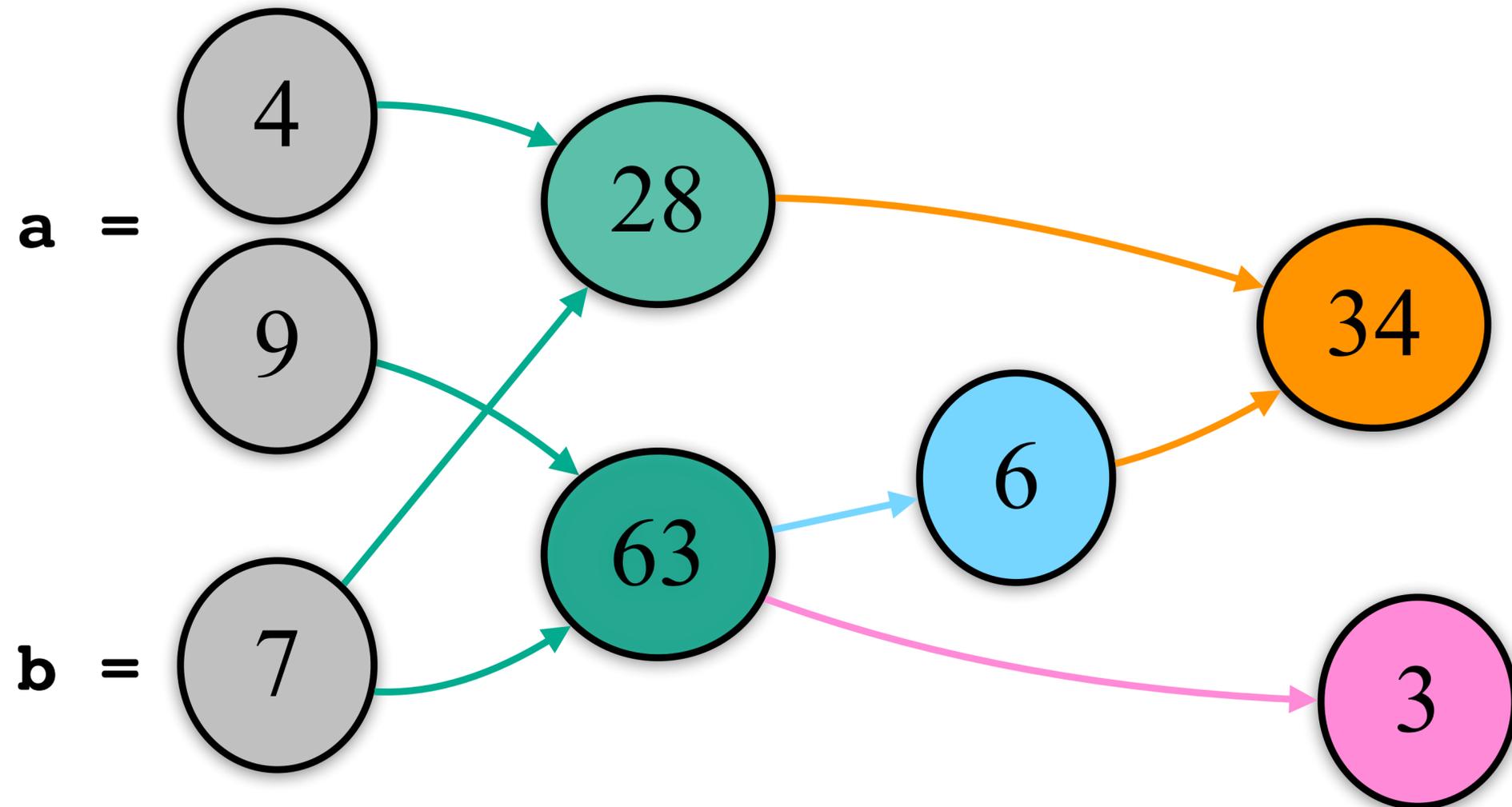
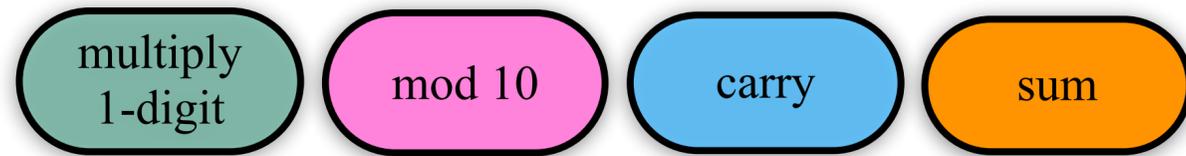
```
function multiply (a[1:p], b[1:q]):  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = a[j] * b[i]  
      t += carry (only if j != p)  
      digits[j] = t mod 10  
      carry = t // 10  
    summands[i] = digits  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] . 10i-1  
  return product
```

Computation graph for 49 x 7



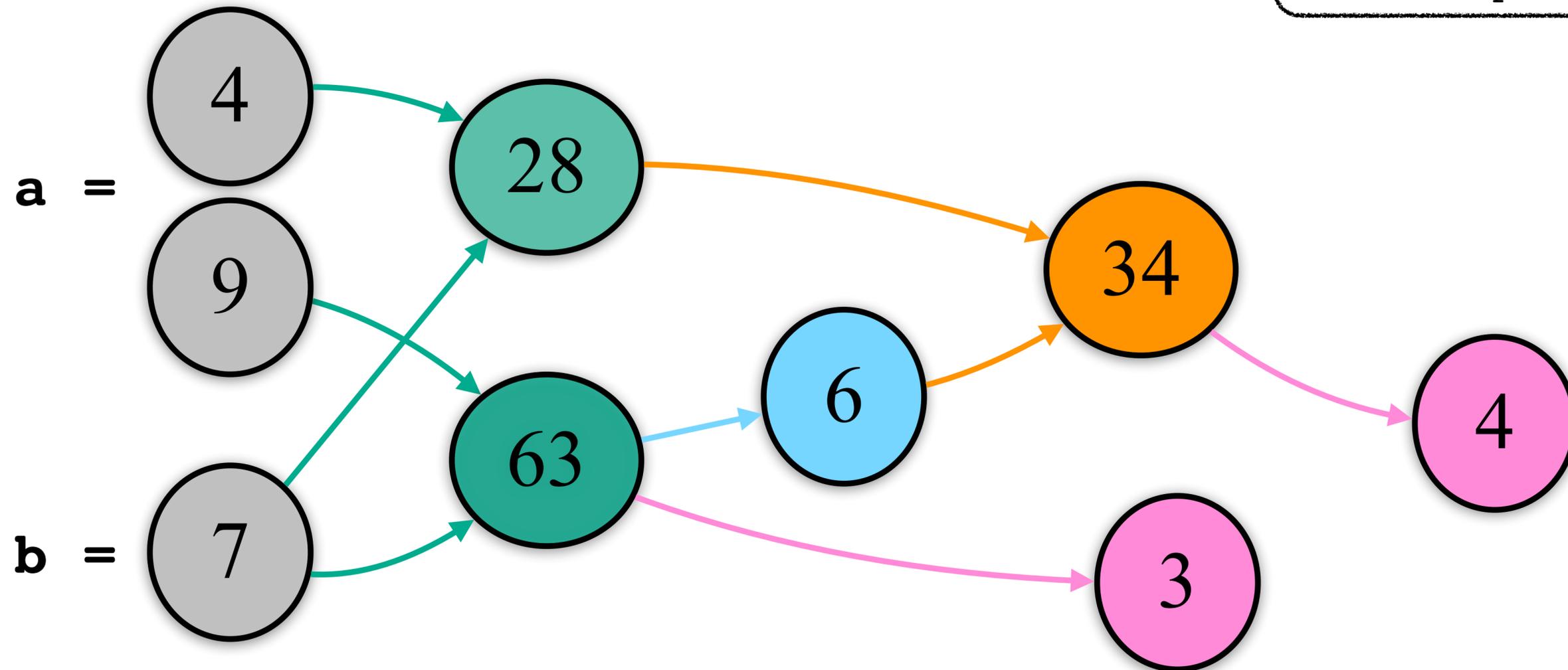
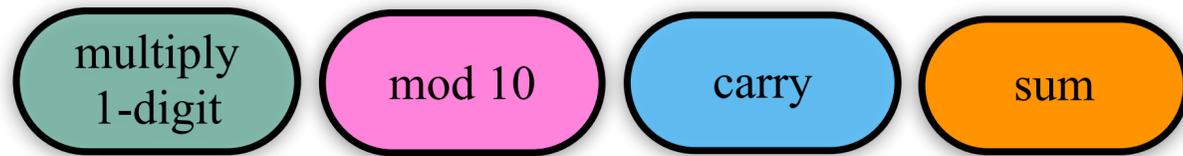
```
function multiply (a[1:p], b[1:q]):  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = a[j] * b[i]  
      t += carry (only if j != p)  
      digits[j] = t mod 10  
      carry = t // 10  
    summands[i] = digits  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] * 10i-1  
  return product
```

Computation graph for 49 x 7



```
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  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] * 10i-1  
  return product
```

Computation graph for 49 x 7



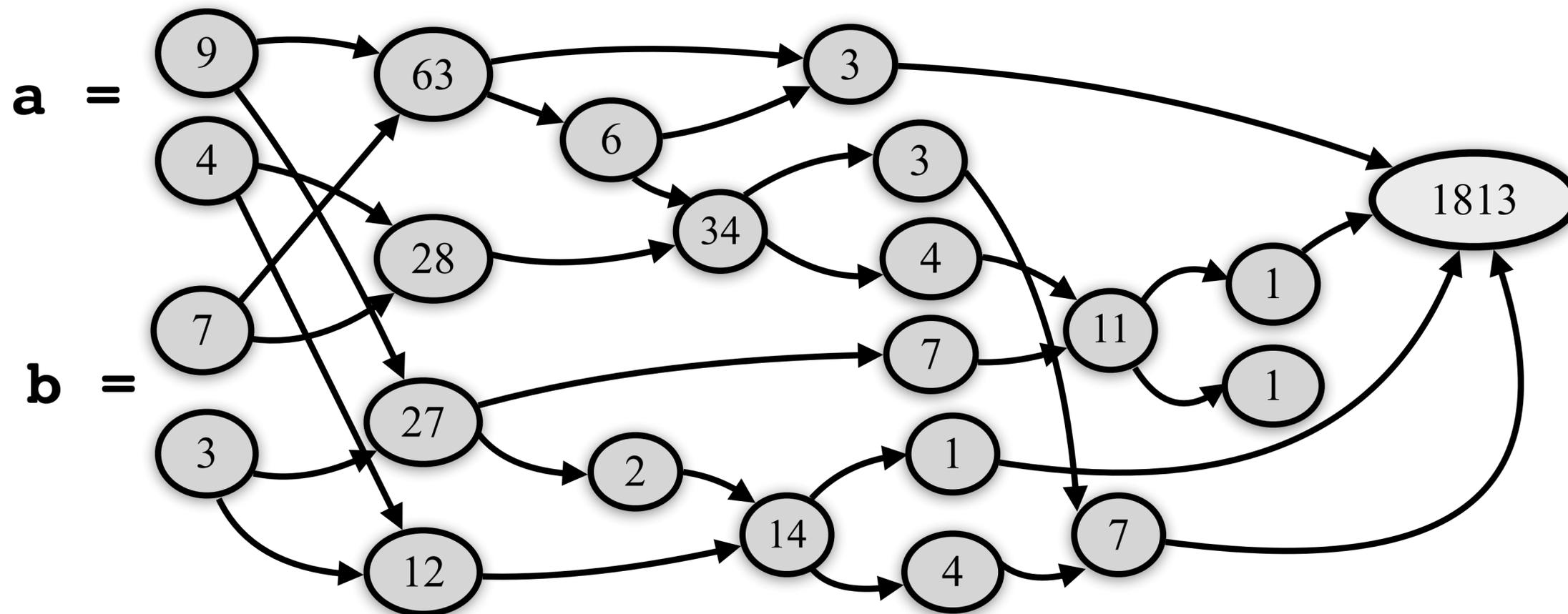
```
function multiply (a[1:p], b[1:q]):  
  for i = q to 1  
    carry = 0  
    for j = p to 1  
      t = a[j] * b[i]  
      t += carry (only if j != p)  
      digits[j] = t mod 10  
      carry = t // 10  
    summands[i] = digits  
  
  product =  $\sum_{i=1}^q$  summands[q+1-i] * 10i-1  
  return 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?*

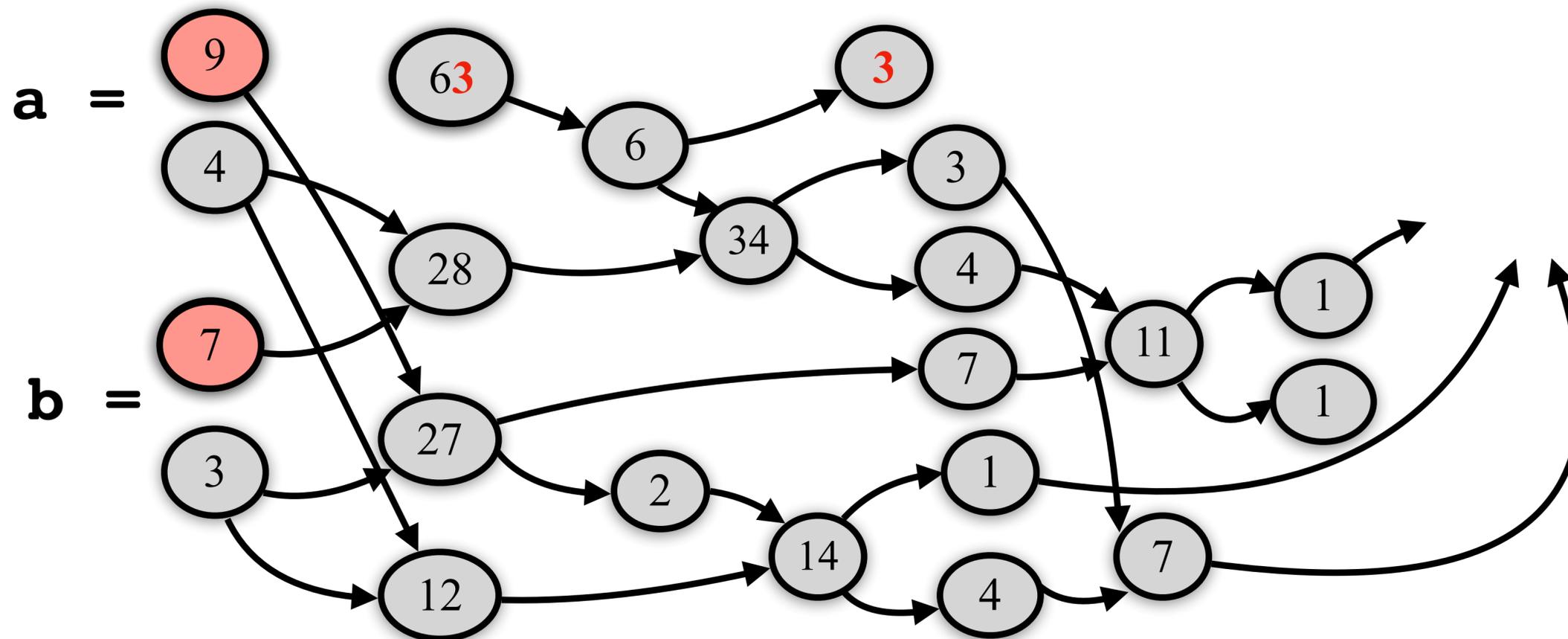
How can we characterize model performance with respect to properties of each task sample?

Information Gain Explains Where Transformers Partially Excel



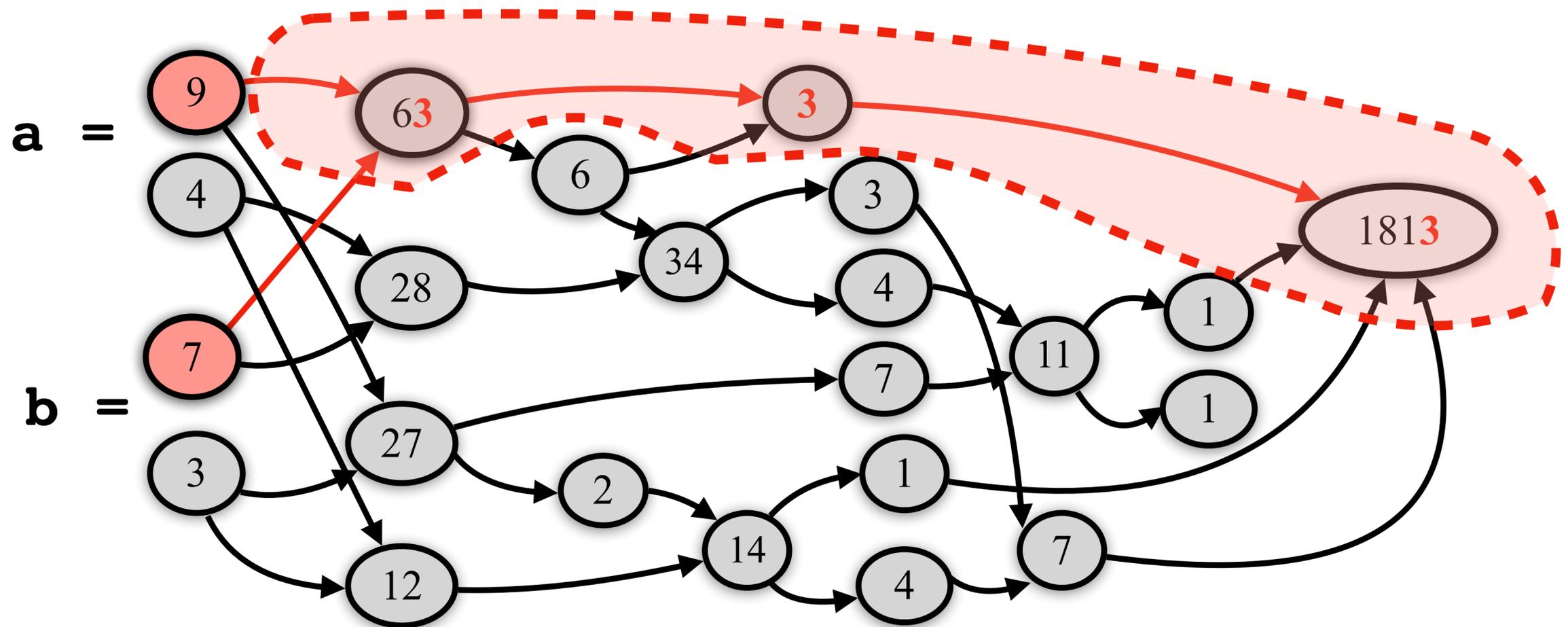
Information Gain Explains Where Transformers Partially Excel

$$\begin{array}{r}
 49 \\
 37 \\
 \hline
 343 \\
 147 \\
 \hline
 = 1813
 \end{array}$$



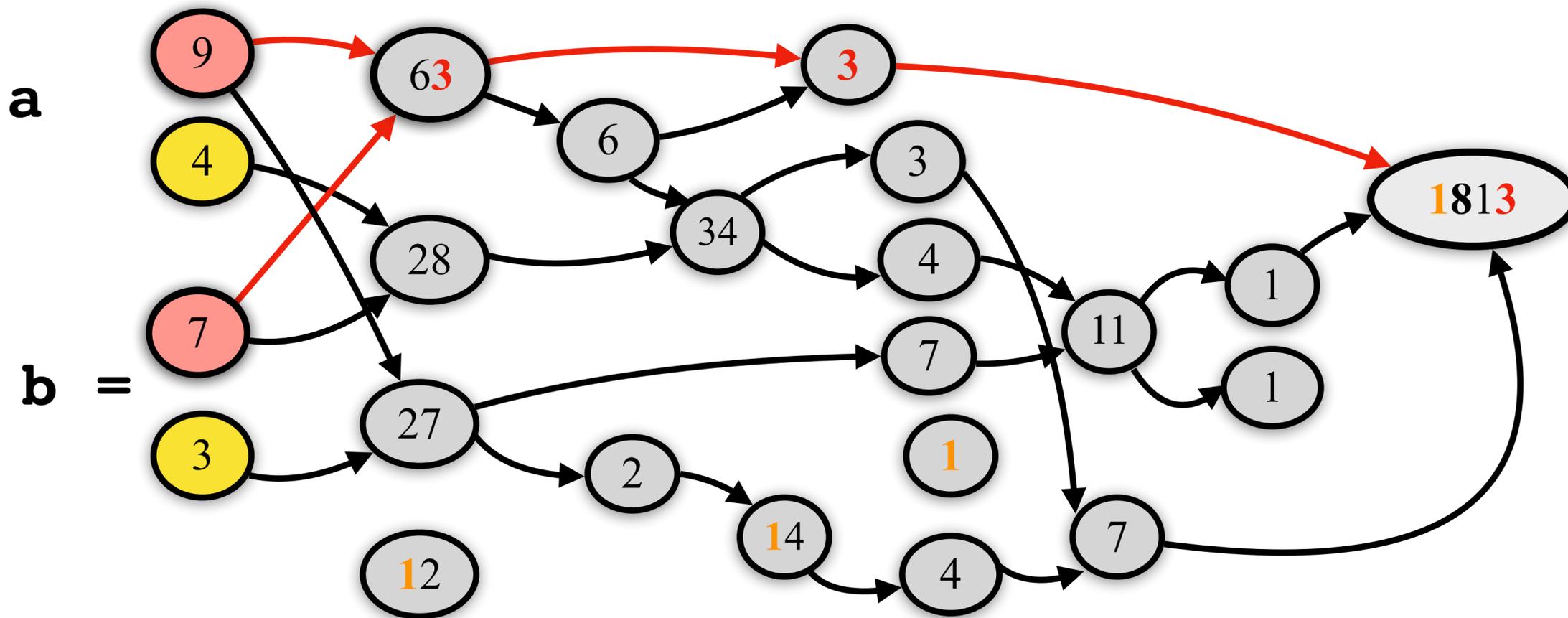
Information Gain Explains Where Transformers Partially Excel

$$\begin{array}{r}
 49 \\
 37 \\
 \hline
 343 \\
 147 \\
 \hline
 = 1813
 \end{array}$$



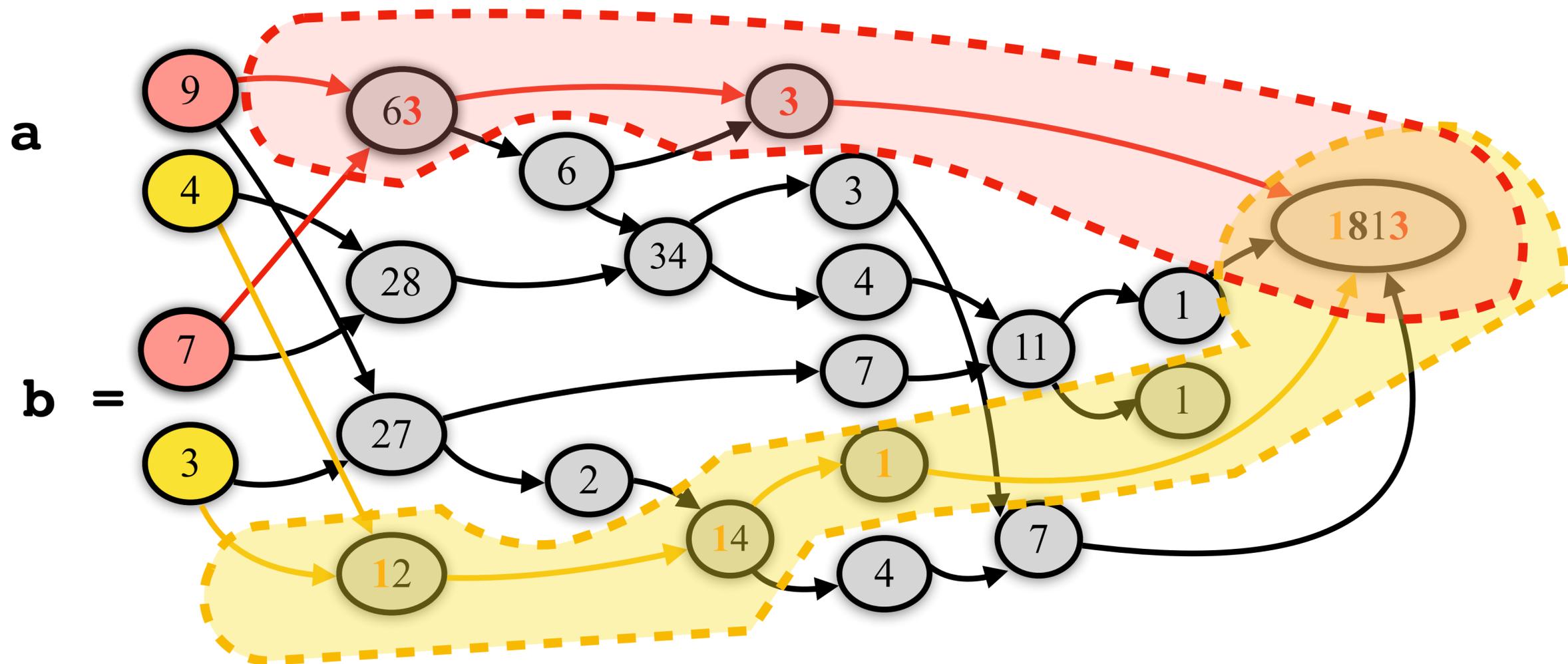
Information Gain Explains Where Transformers Partially Excel

$$\begin{array}{r}
 49 \\
 37 \\
 \hline
 343 \\
 147 \\
 \hline
 = 1813
 \end{array}$$



Information Gain Explains Where Transformers Partially Excel

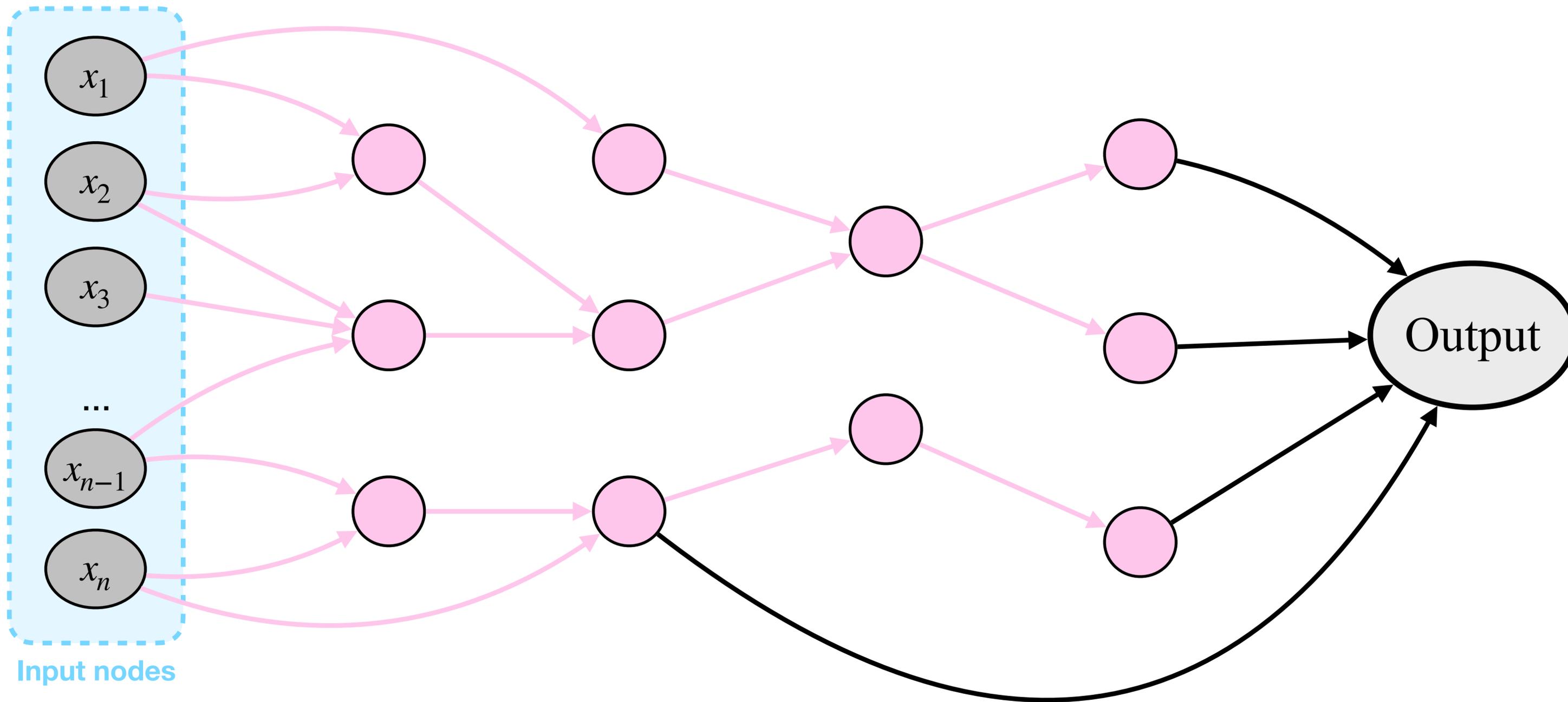
$$\begin{array}{r}
 49 \\
 37 \\
 \hline
 343 \\
 147 \\
 \hline
 = 1813
 \end{array}$$

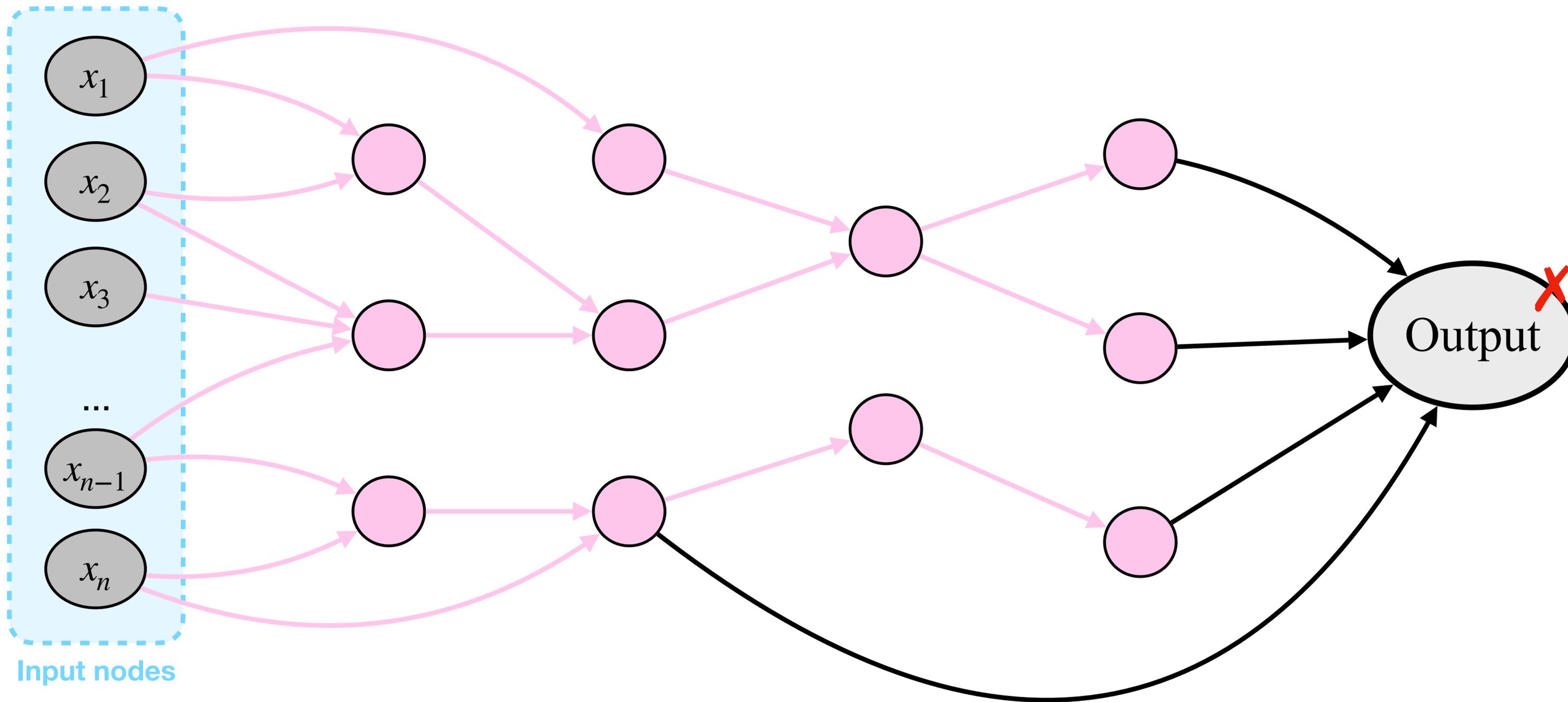


How can we characterize model performance with respect to properties of each task sample?

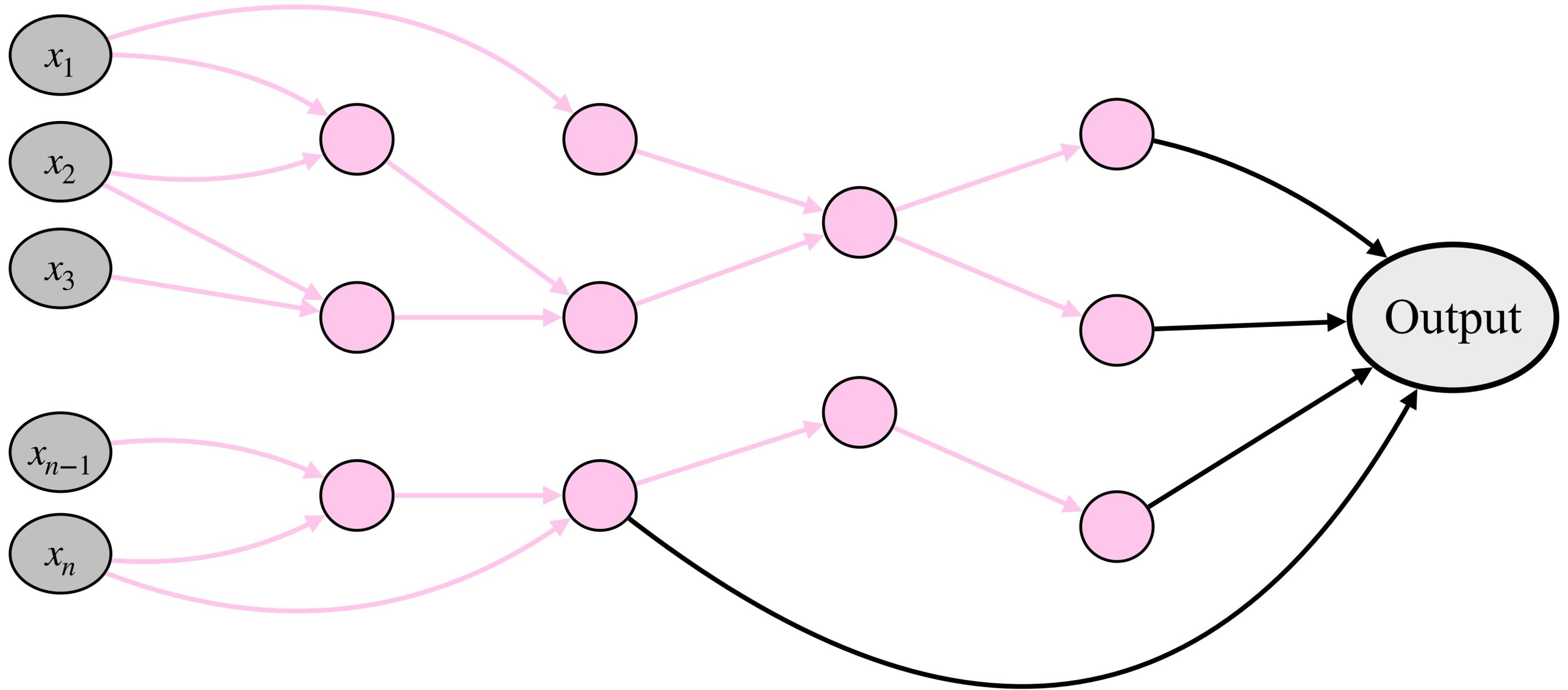
*When we see models solving a seemingly difficult question, what can we correlate it to?
How do we characterize model errors?*

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

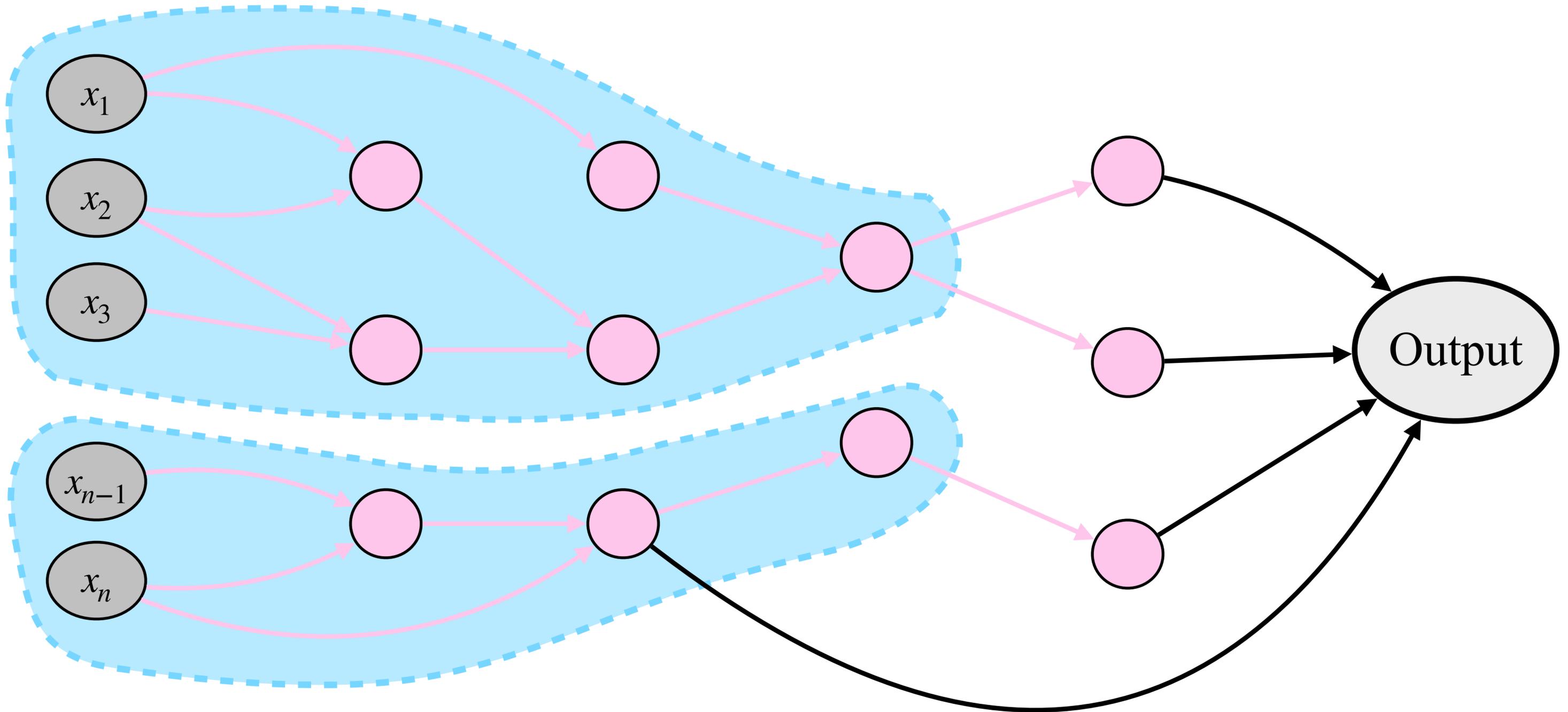




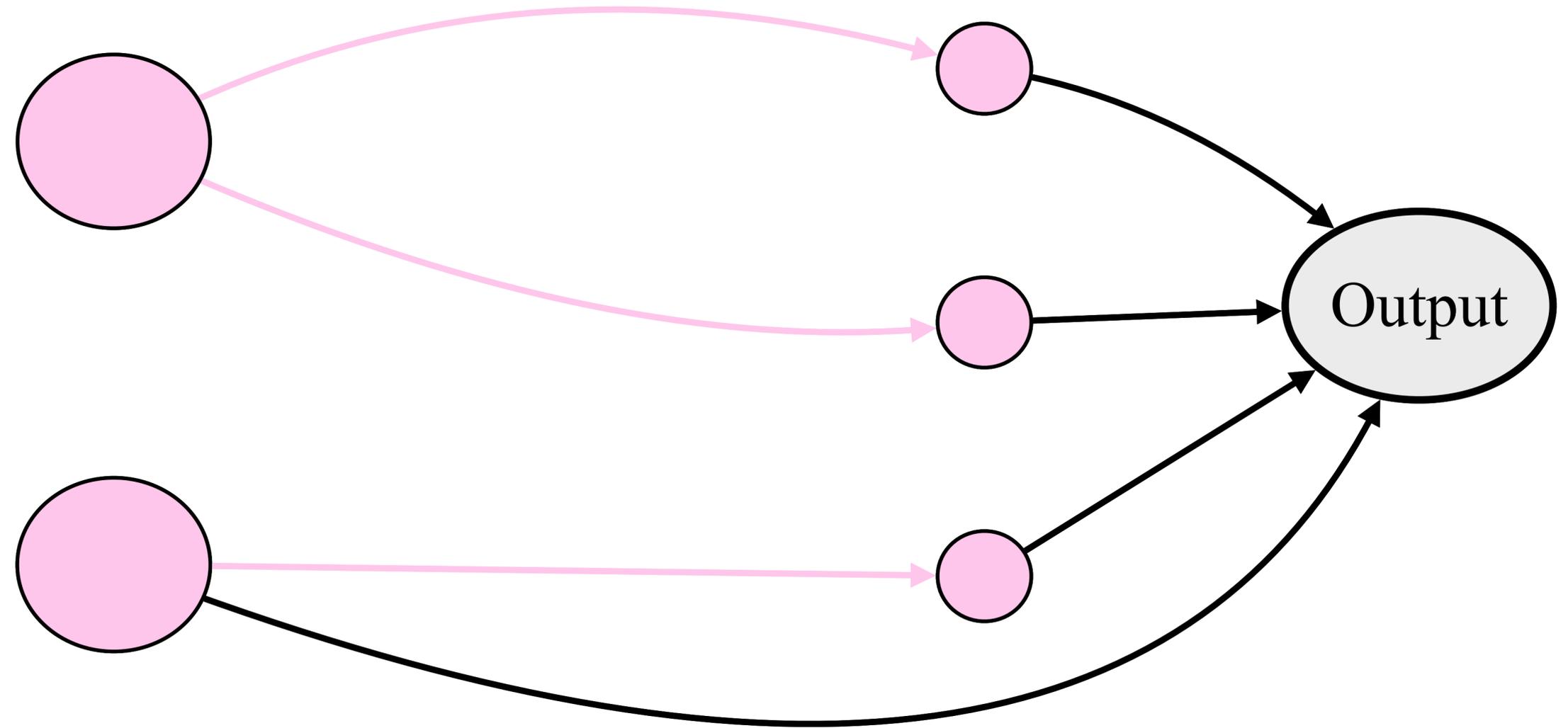
What is the correlation between a model generating a correct output and having seen relevant subgraphs during training?

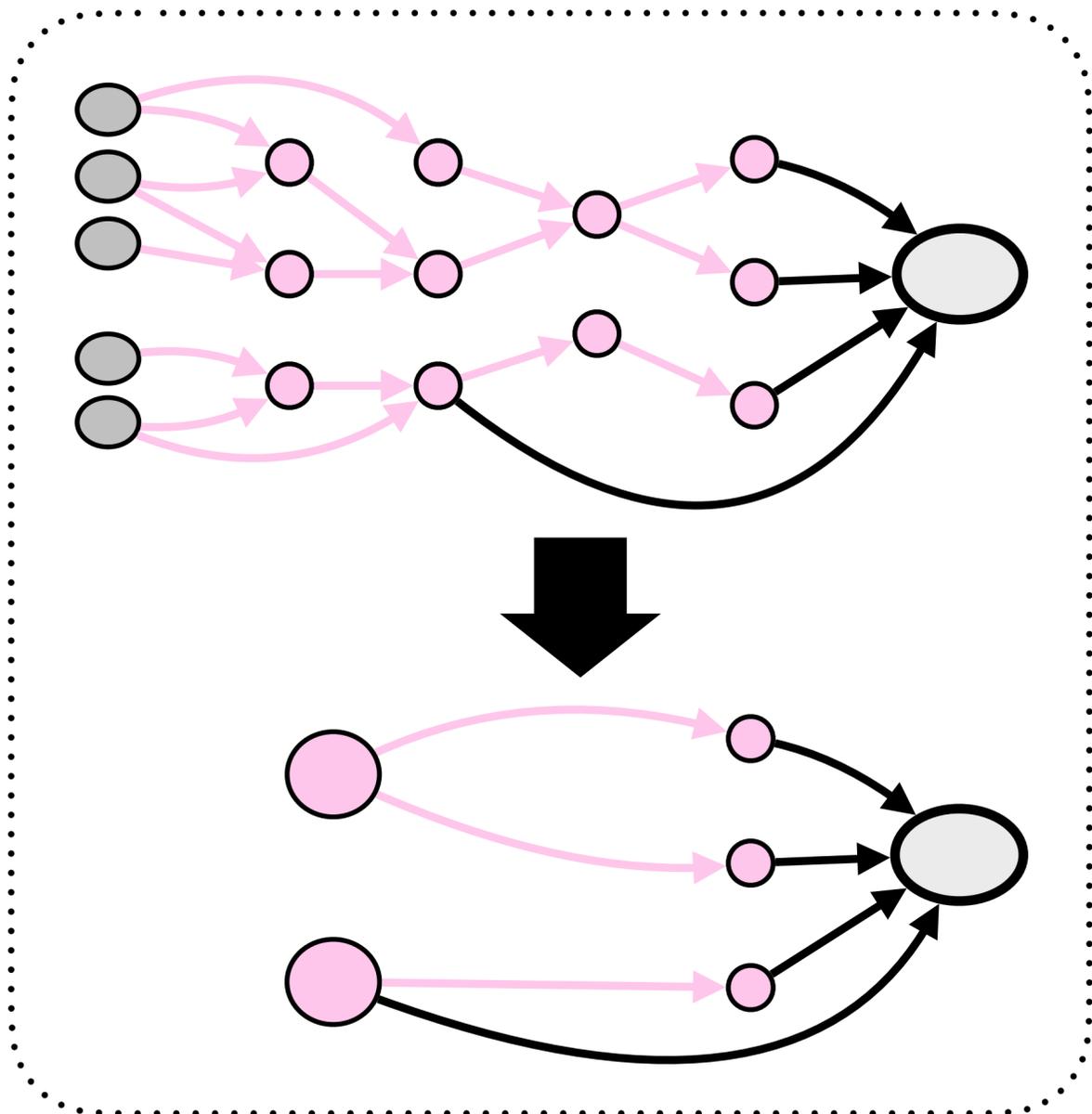


Detect subgraphs already seen during training:

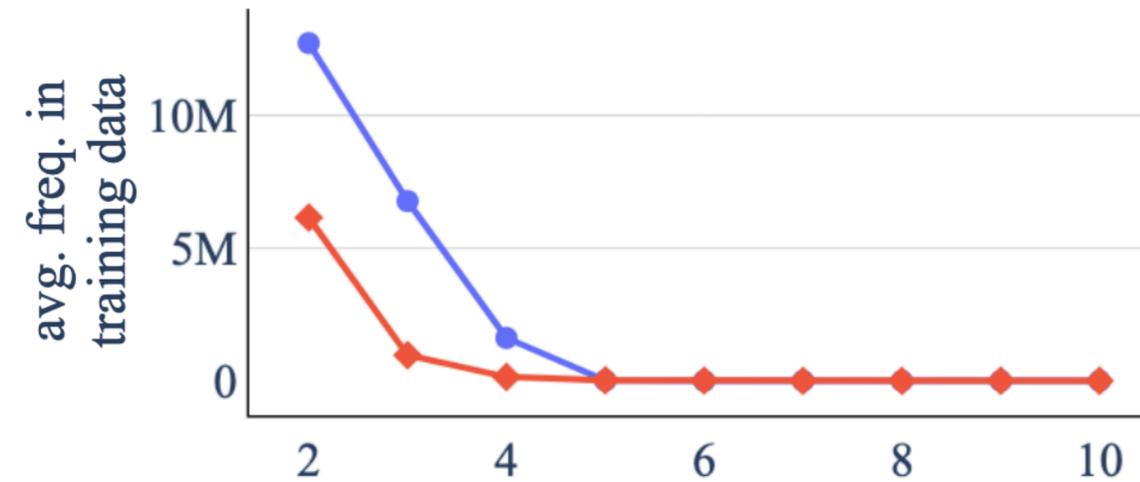


*If you already saw relevant subgraphs during training, the inference is only **seemingly** highly compositional*

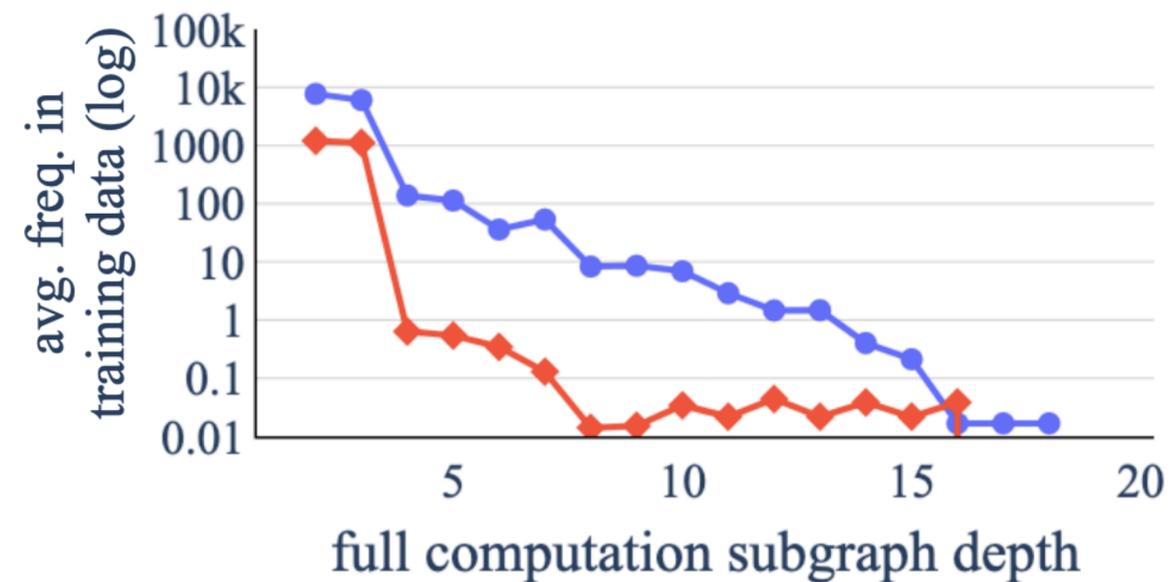




Fine-tuned GPT3 - Multiplication

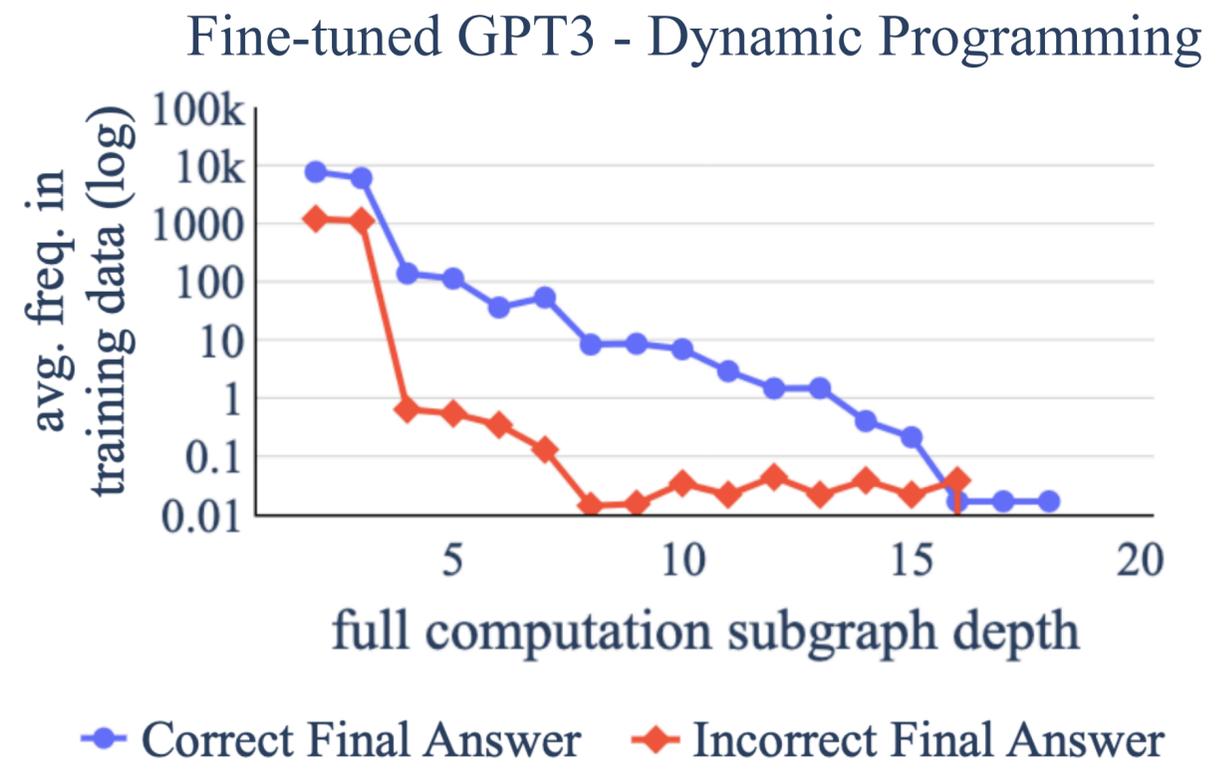
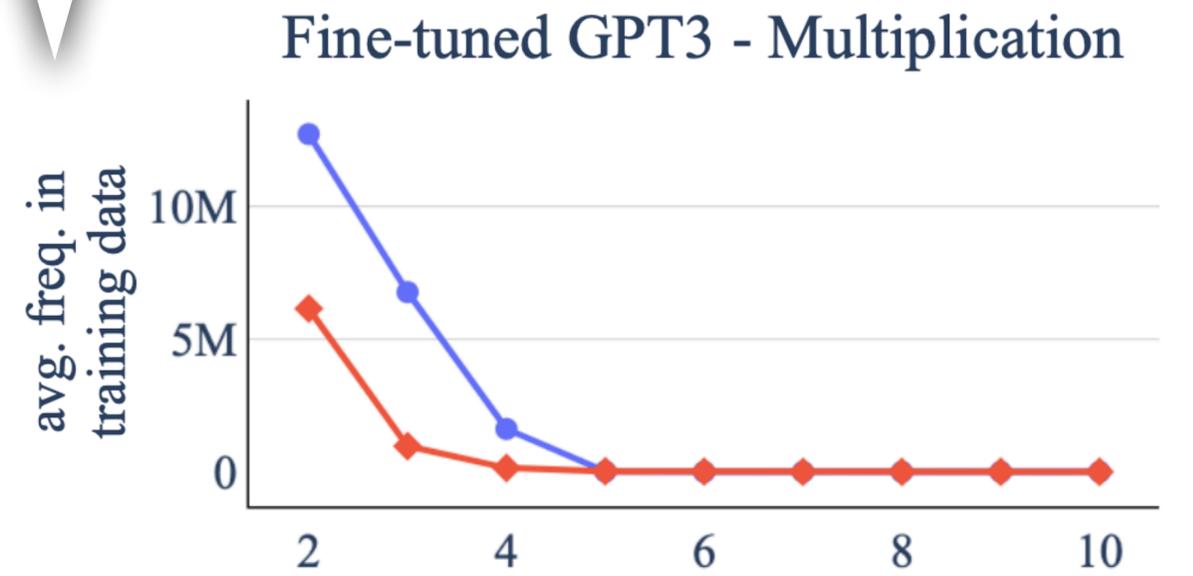
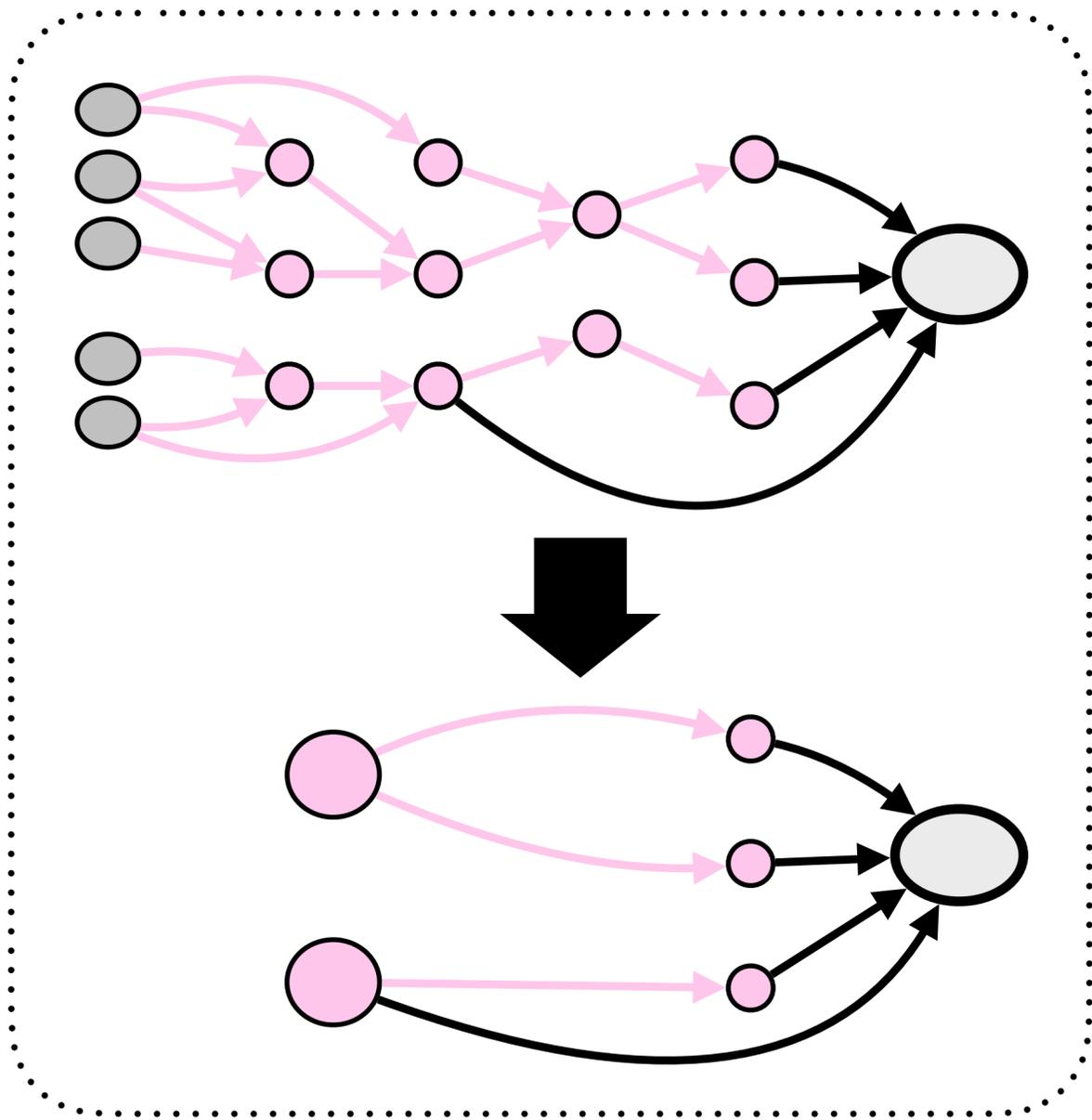


Fine-tuned GPT3 - Dynamic Programming



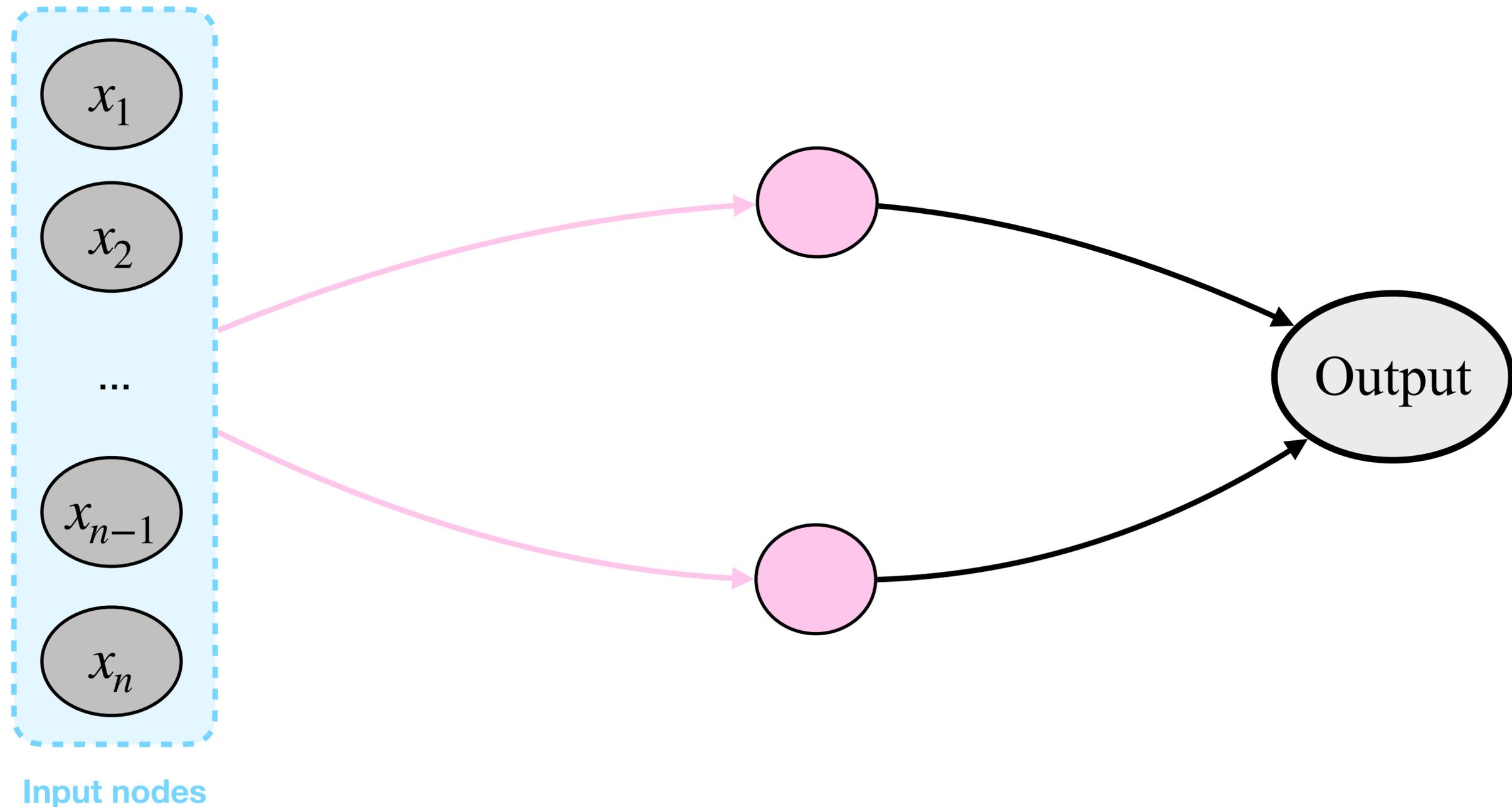
● Correct Final Answer ◆ Incorrect Final Answer

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

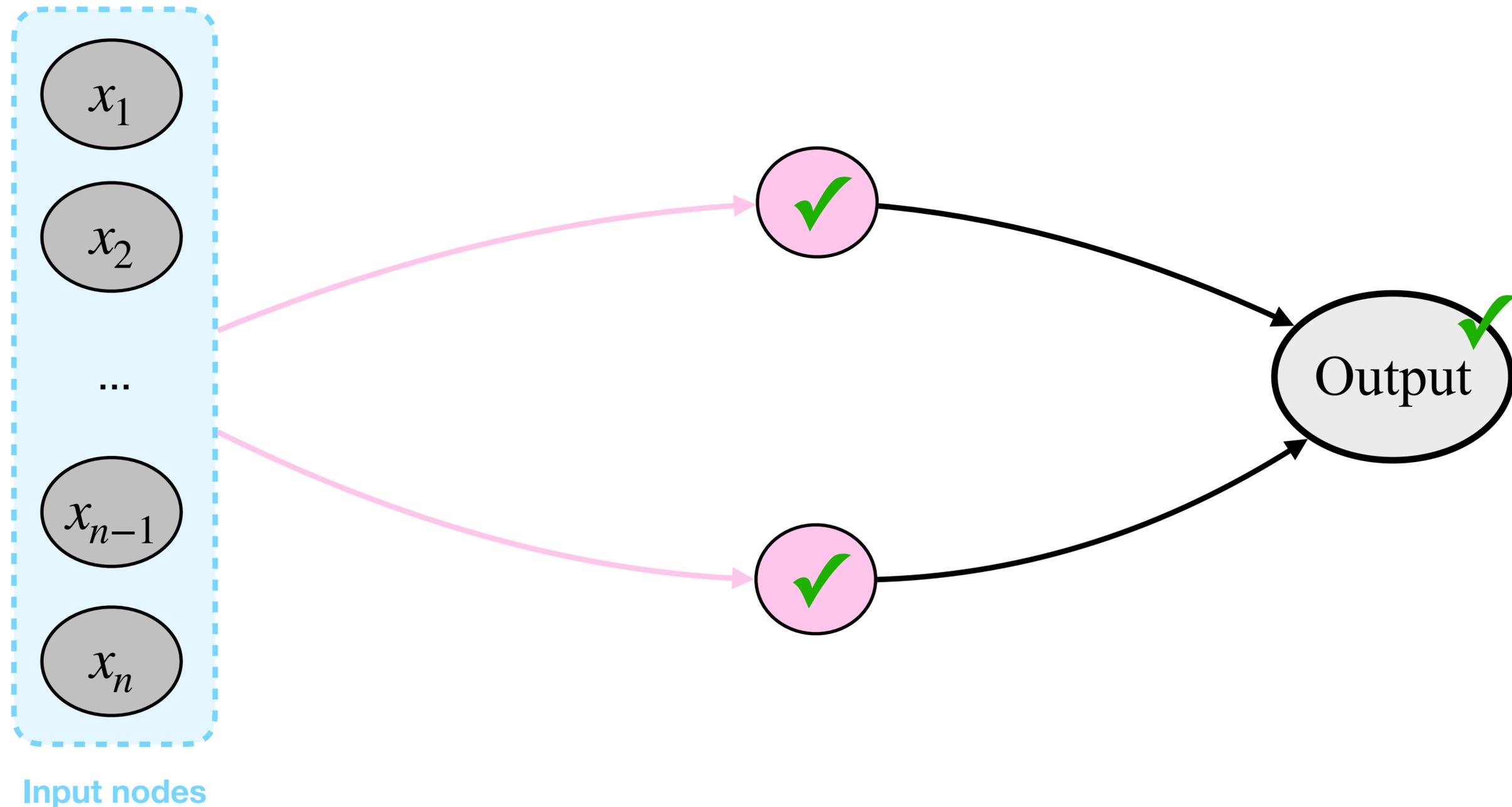


**Why does performance inevitably
decay with problem size?**

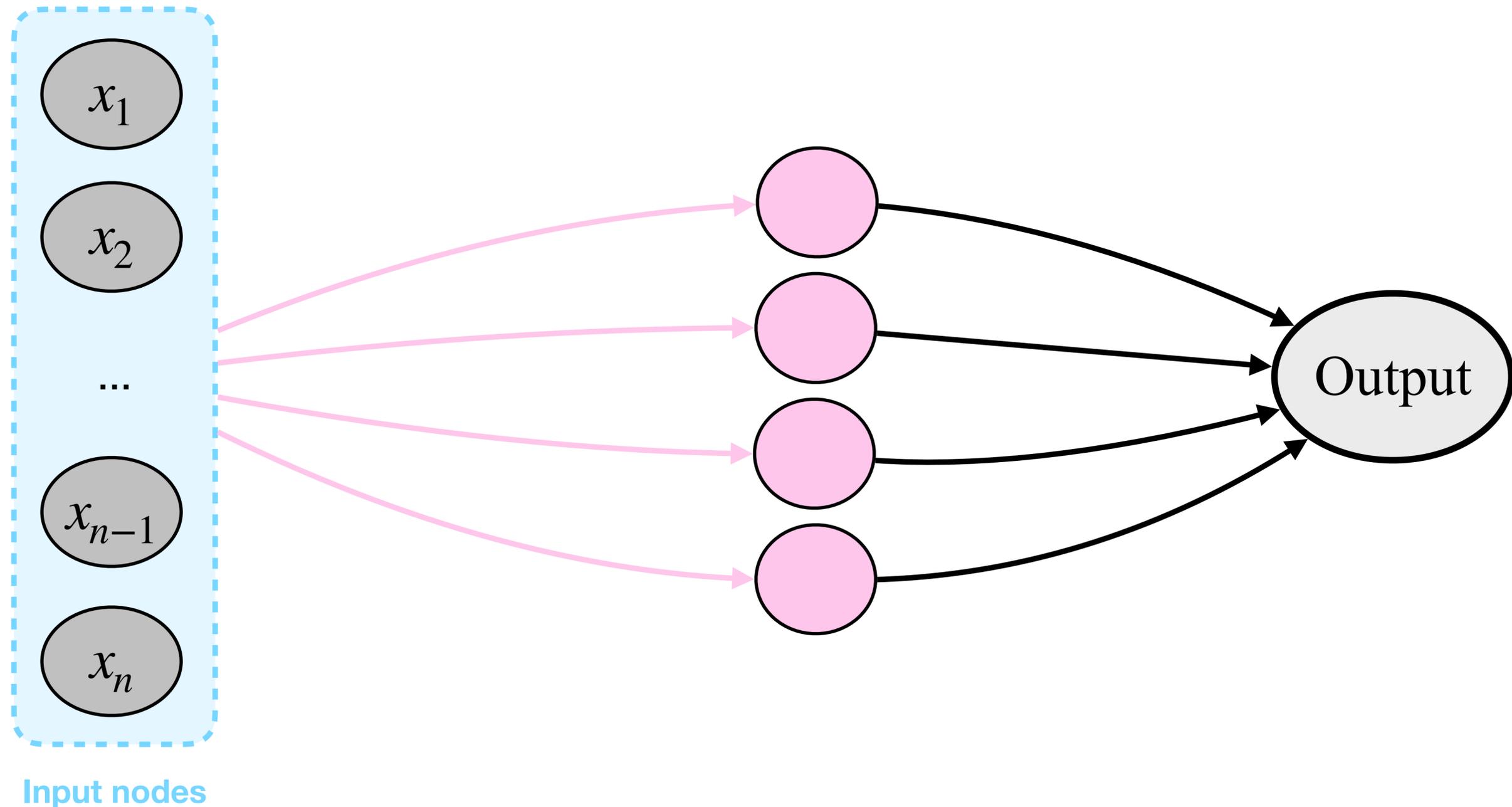
An increasing number of **independent** reasoning steps inevitably lead to errors



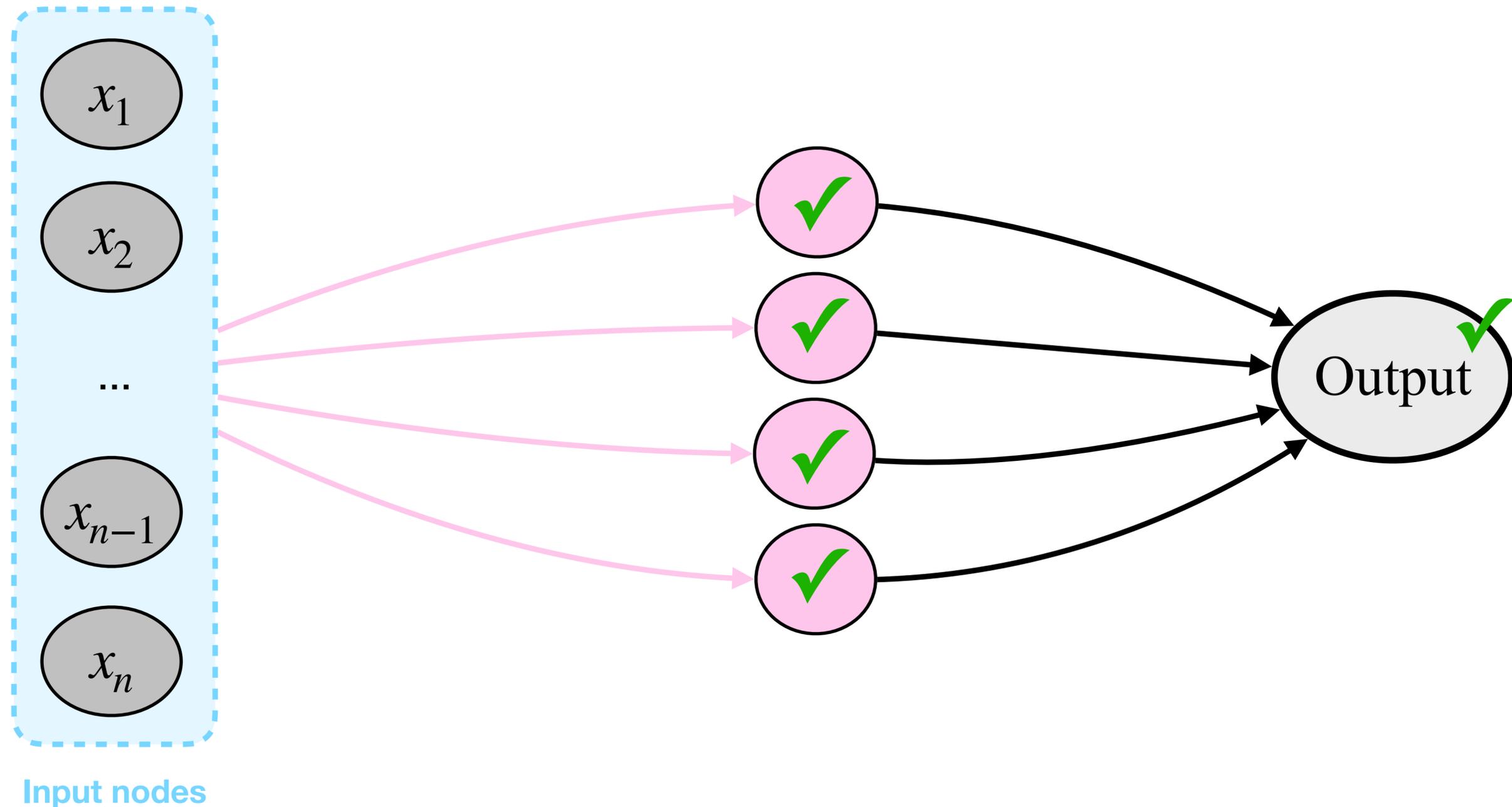
An increasing number of **independent** reasoning steps inevitably lead to errors



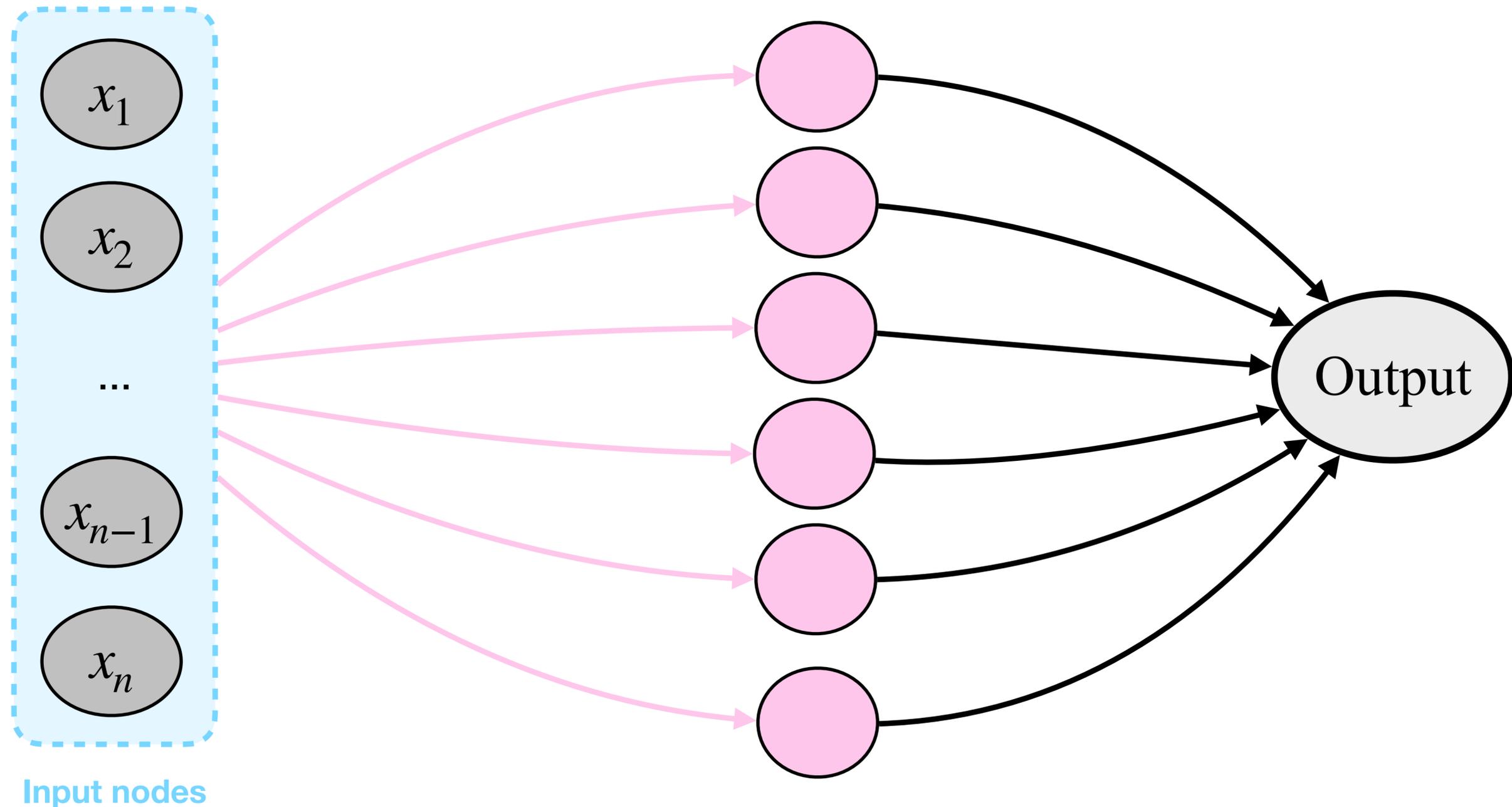
An increasing number of **independent** reasoning steps inevitably lead to errors



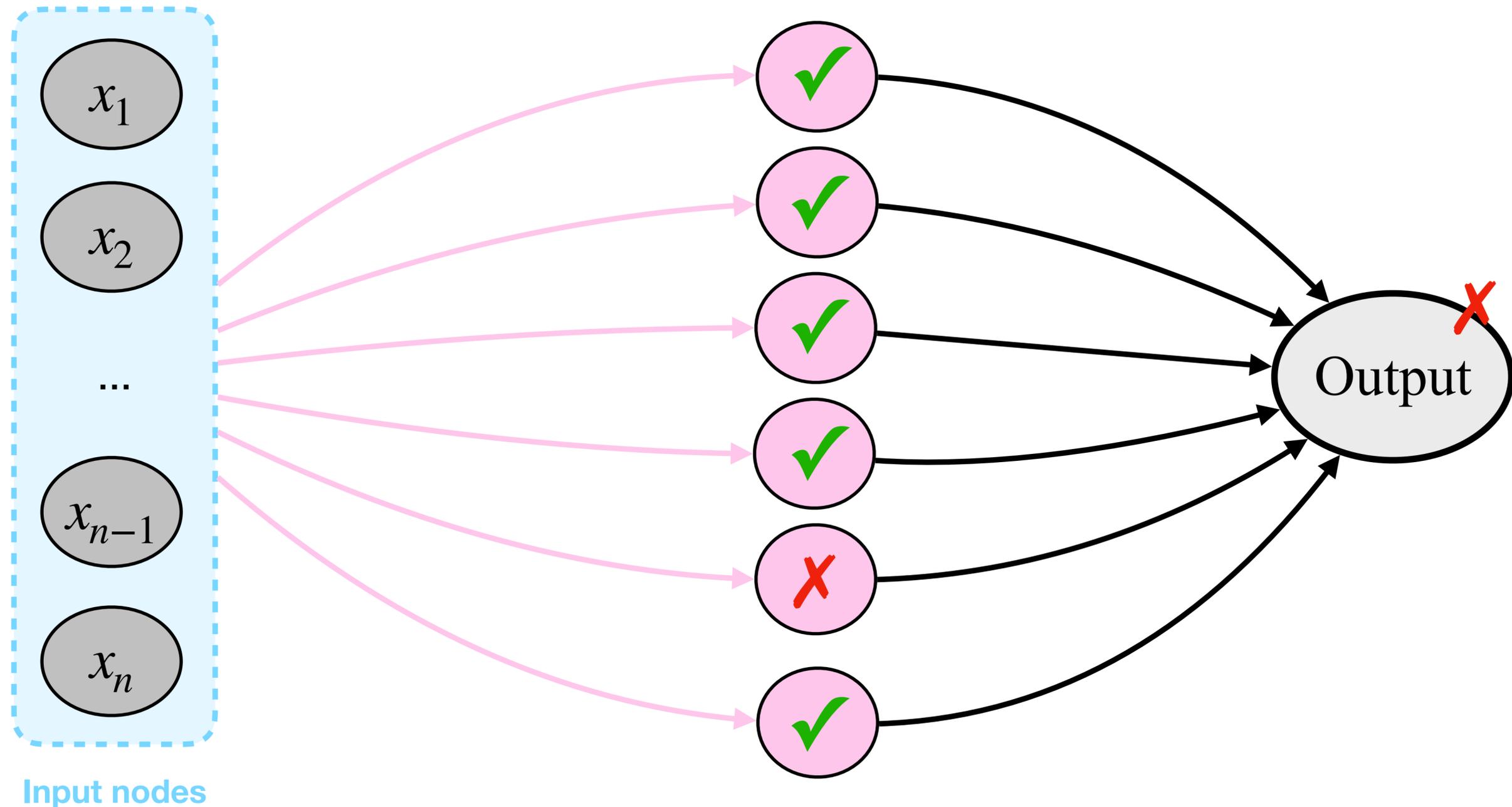
An increasing number of **independent** reasoning steps inevitably lead to errors



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

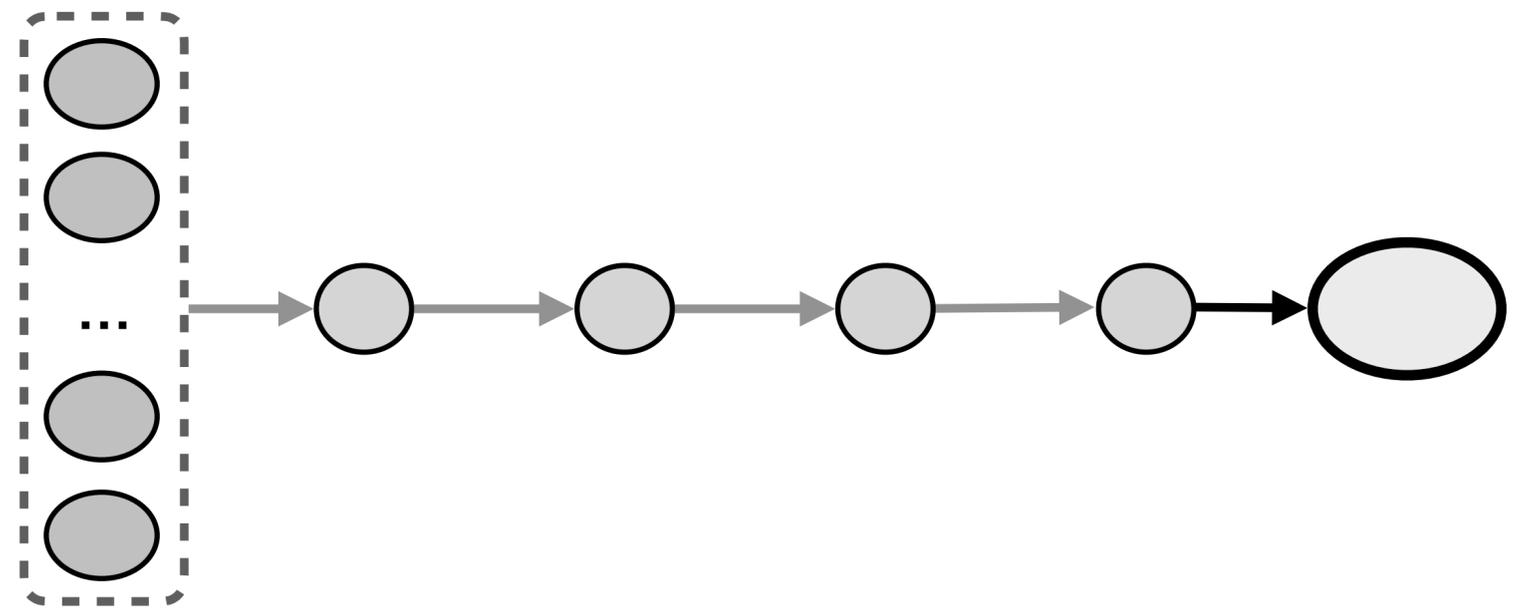
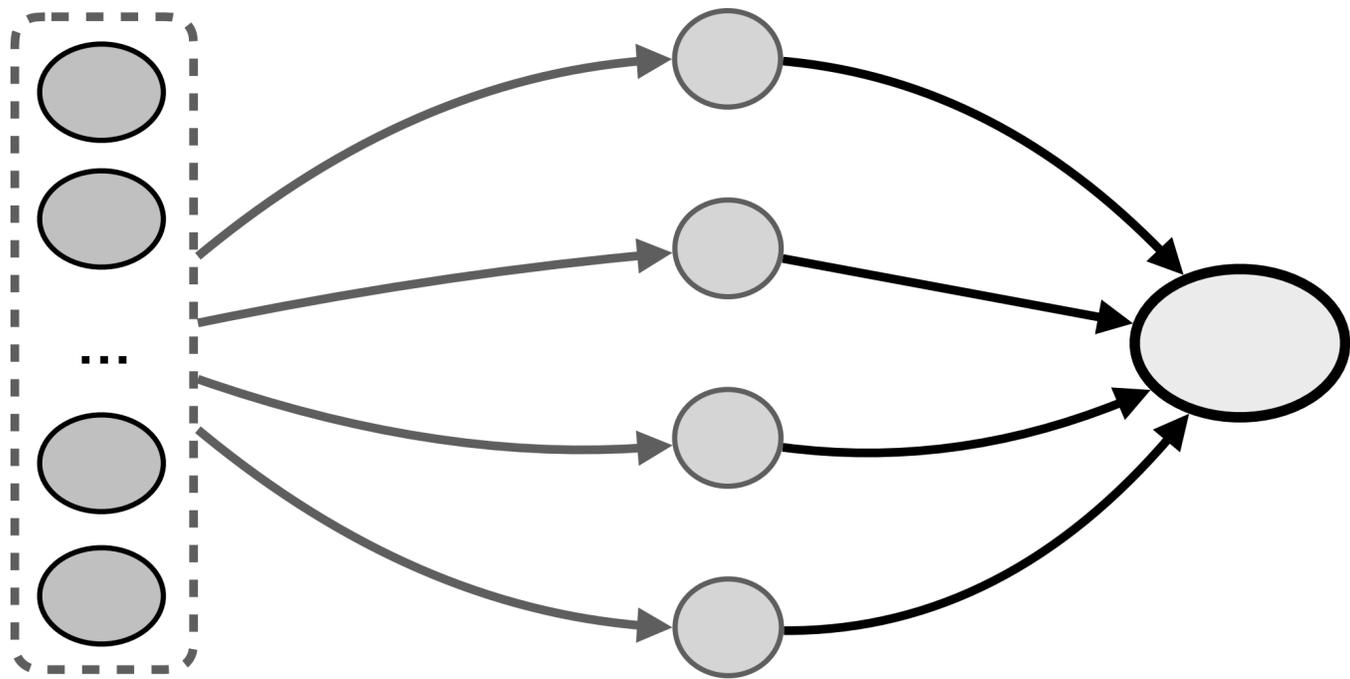


An increasing number of **independent** reasoning steps inevitably lead to errors



An increasing number of reasoning steps inevitably lead to errors

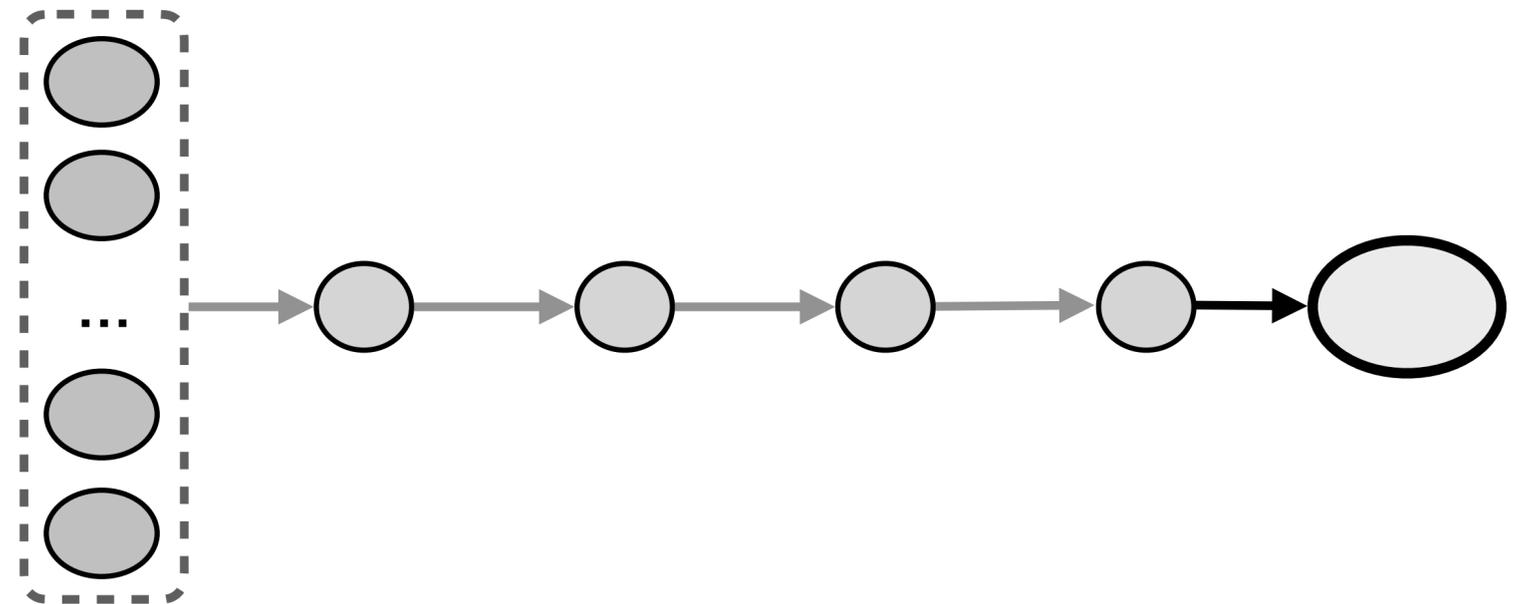
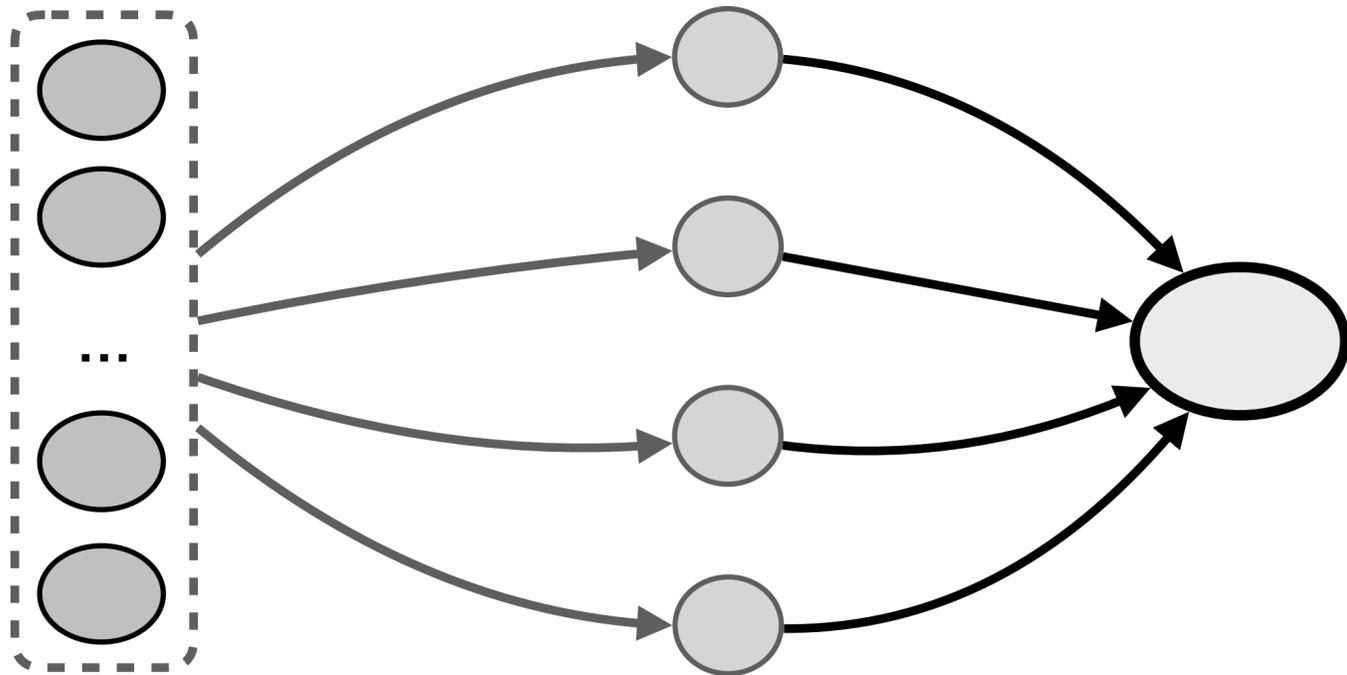
If the probability of making an error in a single reasoning step is ϵ , probability of success is...



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

with exponentially increasing probability

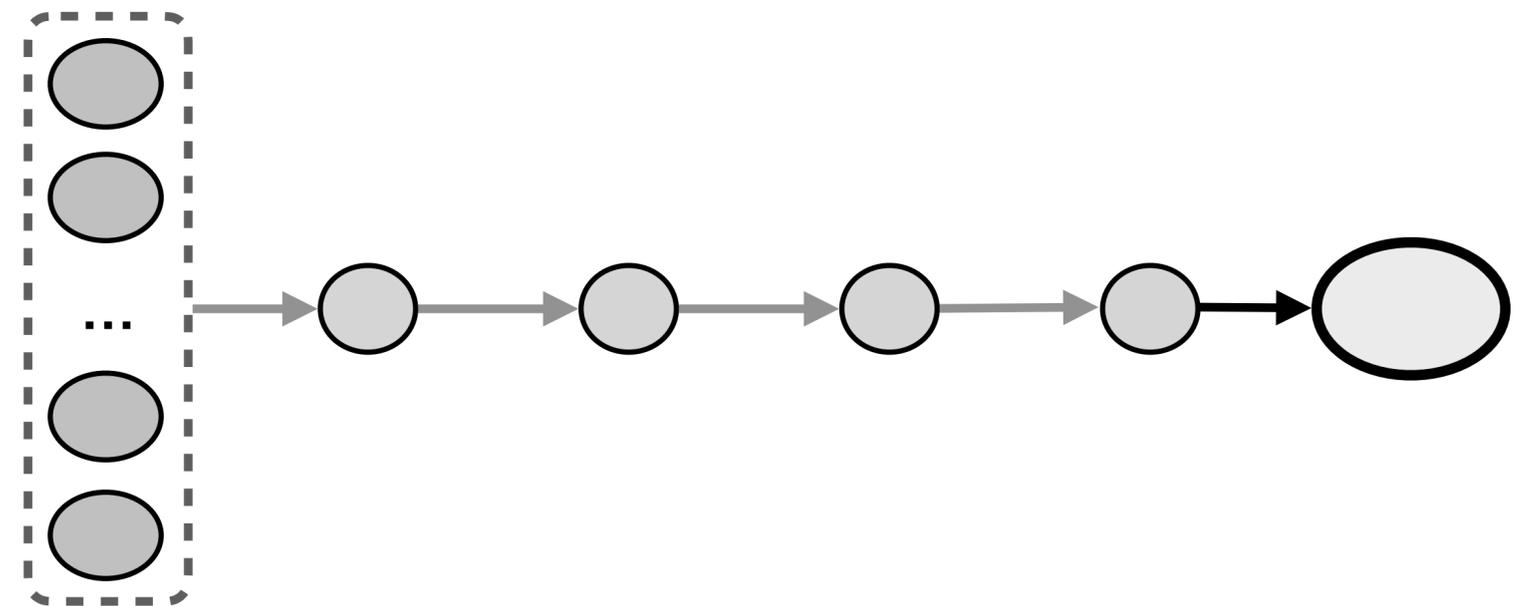
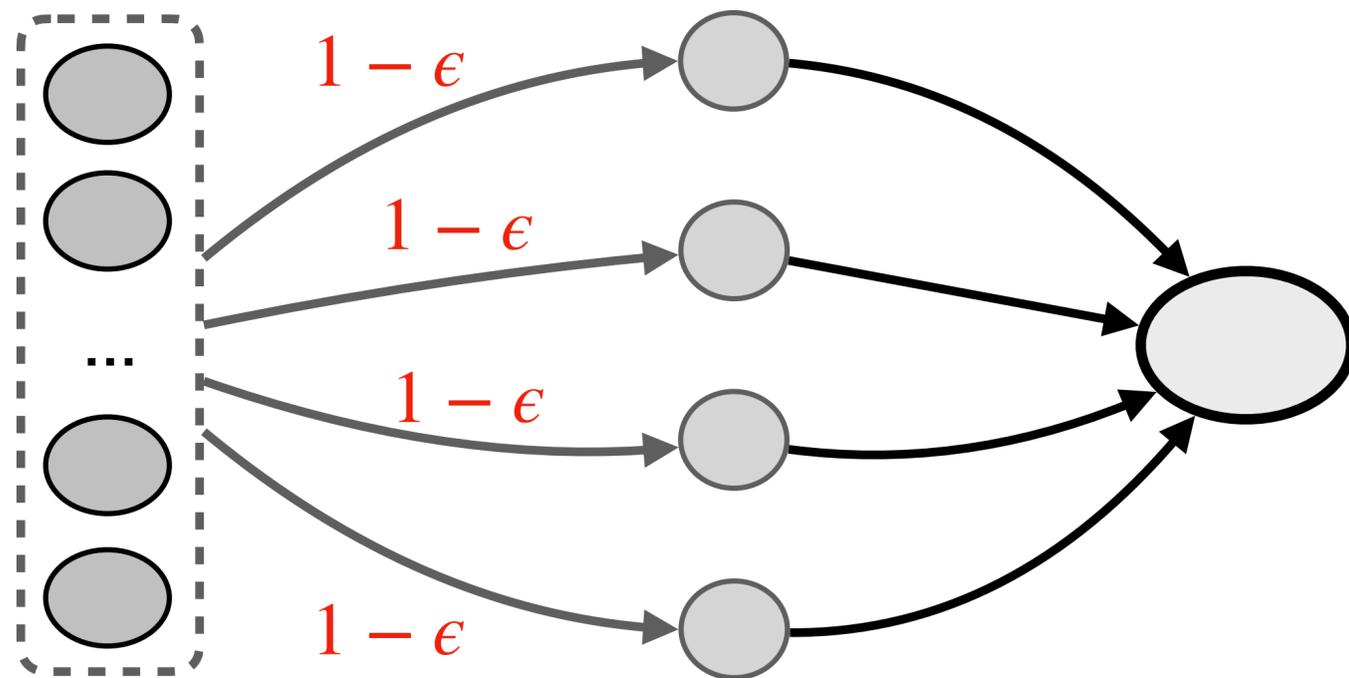
If the probability of making an error in a single reasoning step is ϵ , probability of success is...



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...

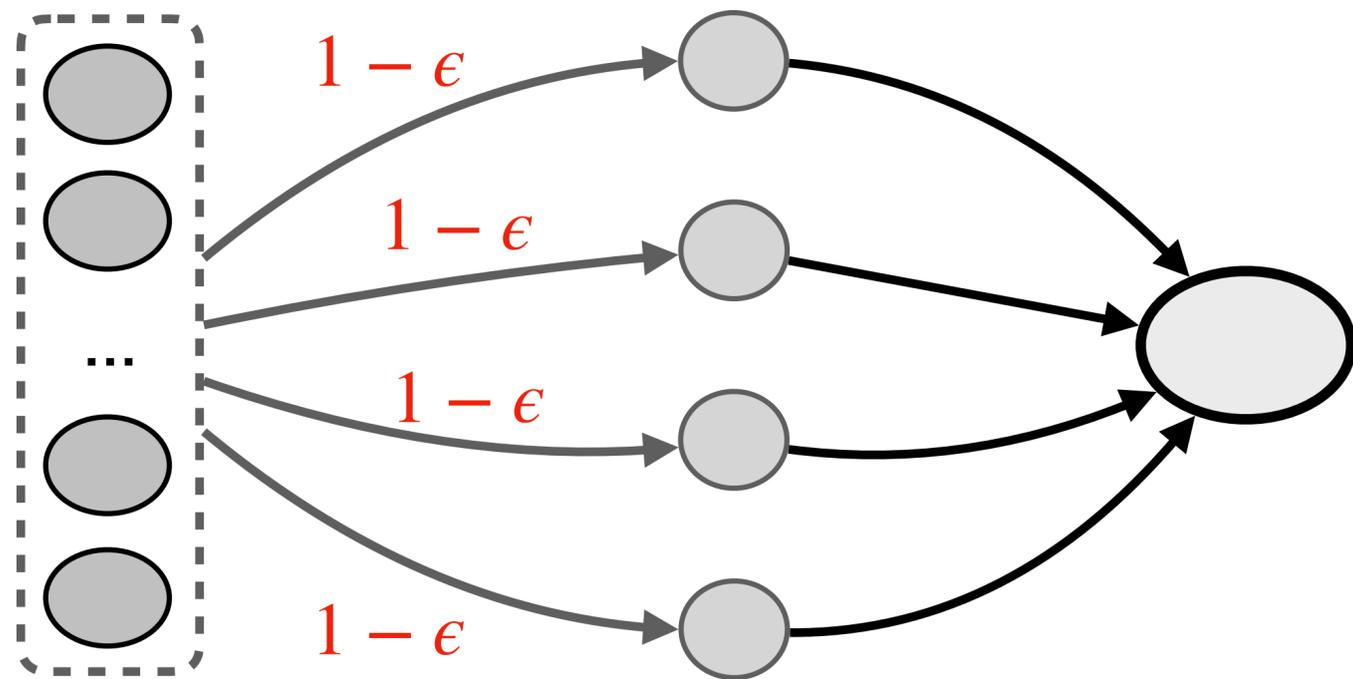


$$\approx (1 - \epsilon)^n$$

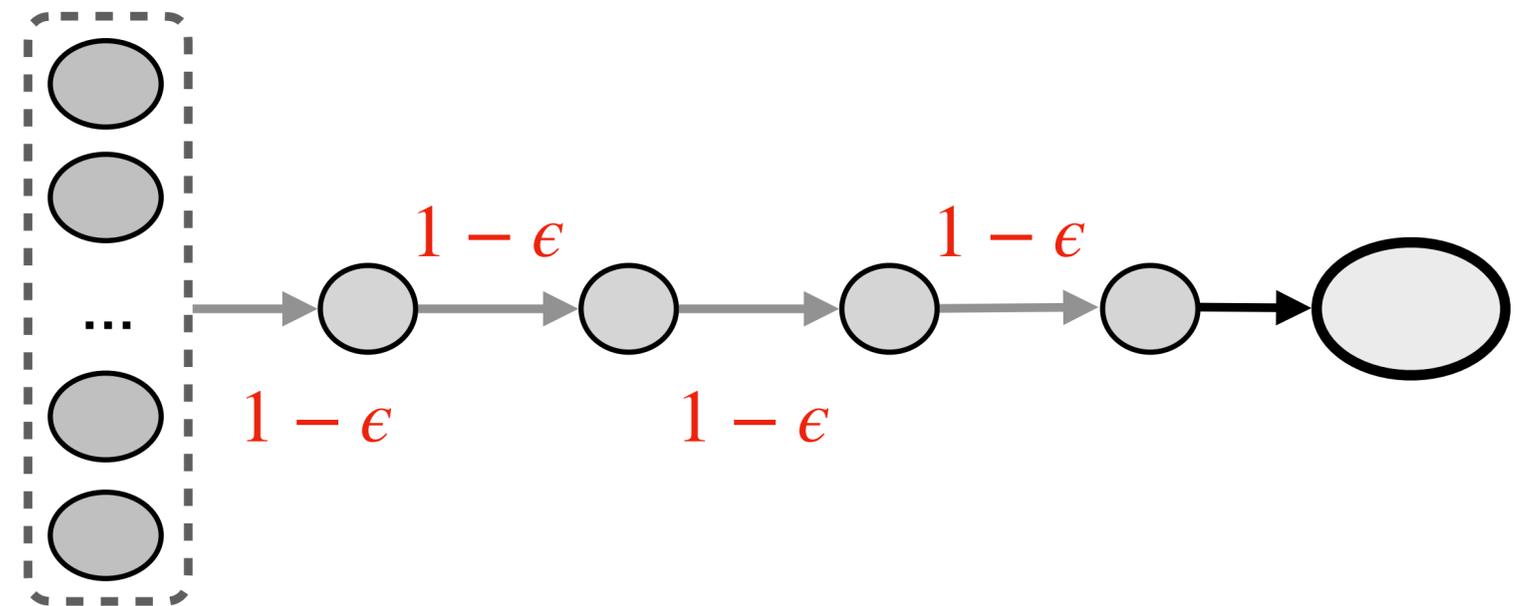
An increasing number of reasoning steps inevitably ~~lead to errors~~

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

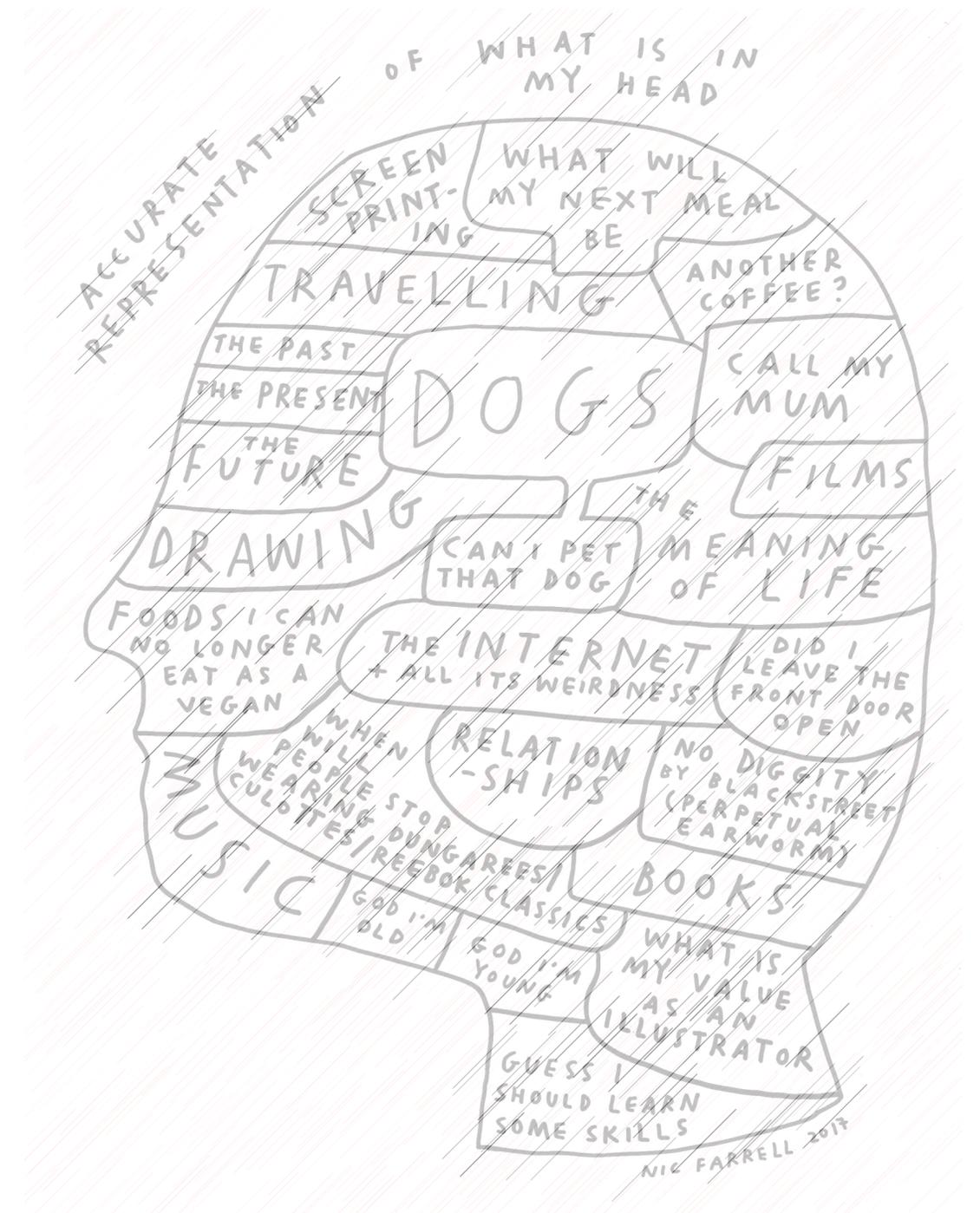
- Does it mean there aren't avenues for progress? No!
Promising avenues:
 - use transformers in ways that **chain only few compositional steps** to reach a solution
 - use transformers in tasks where **evaluation metrics afford leniency**
 - **augmenting transformers with planning modules** and refining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



GPT-4 already showed theory of mind?

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.**"*

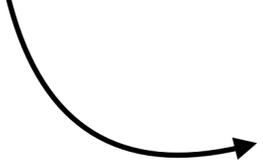
GPT-4 already showed theory of mind?

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 showed theory of mind?

ChatGPT 4 ▾

 **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:

¹Stanford University, Stanford, CA94305, USA

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?

ChatGPT 4 ▾

Even a slightest change in the label with an unseen word causes GPT-4 to fail

 **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?

How can we systematically quantify theory of mind reasoning skills?

Explore Theory of Mind: program-guided adversarial data generation for theory of mind reasoning

ICLR 2025

W



*Melanie
Sclar*



*Jane
Dwivedi-Yu Fazel-Zarandi*



Maryam



*Yulia
Tsvetkov*



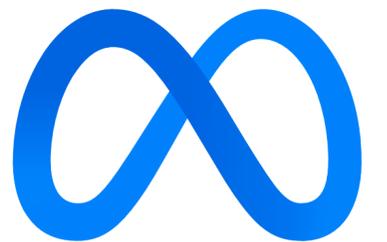
*Yonatan
Bisk*



*Yejin
Choi*



*Asli
Celikyilmaz*

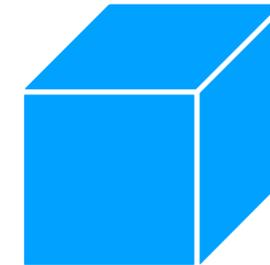
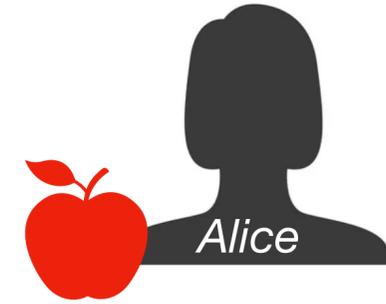
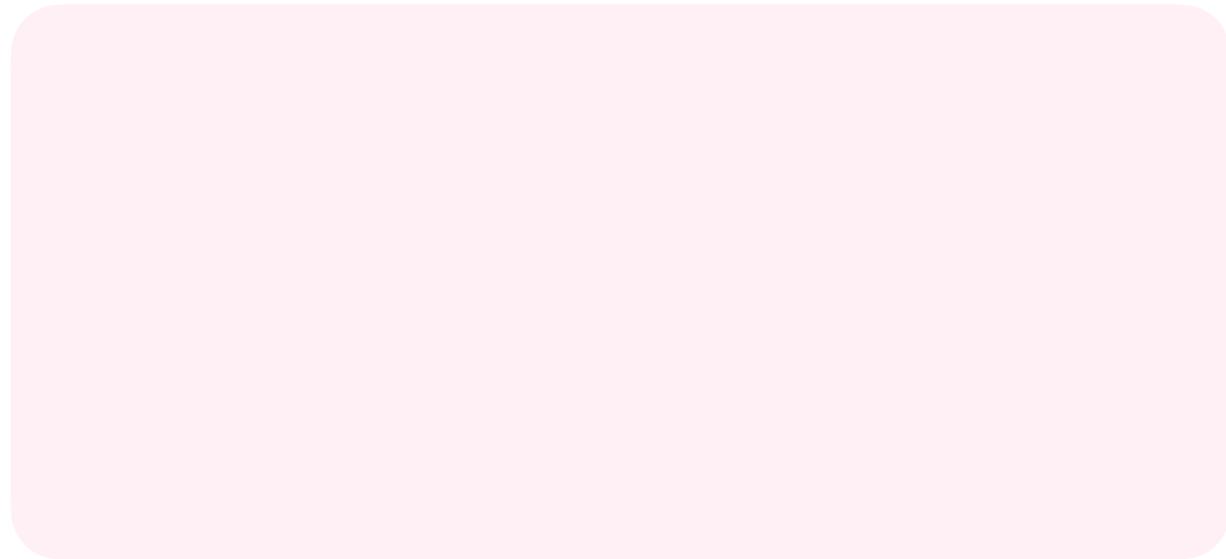


Motivation

- 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.**

Sally-Anne Test

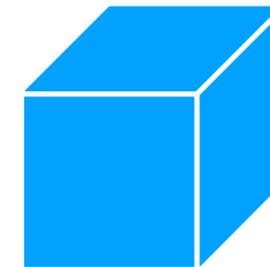
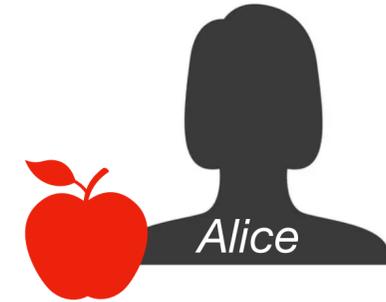
(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)



Sally-Anne Test

(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

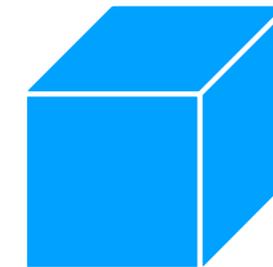


Sally-Anne Test

(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

*Alice puts an apple in the basket,
and then leaves the room.*



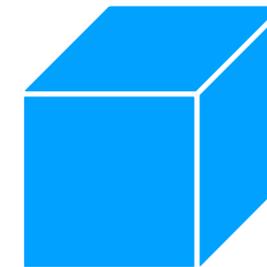
Sally-Anne Test

(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

*Alice puts an apple in the basket,
and then leaves the room.*

Bob then moves the apple to the box.



Sally-Anne Test

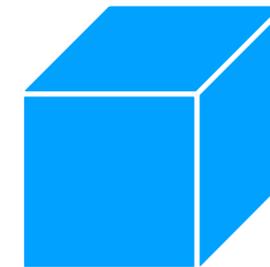
(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

*Alice puts an apple in the basket,
and then leaves the room.*

Bob then moves the apple to the box.

- 1. Where will Bob search for the apple?** *Answer: Box*
- 2. Where does Bob think that Alice will search for the apple when she comes back?** *Answer: Basket*



Sally-Anne Test

(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

*Alice puts an apple in the basket,
and then leaves the room.*

Bob then moves the apple to the box.

- 1. Where will Bob search for the apple?** *Answer: Box*
- 2. Where does Bob think that Alice will search for the apple when she comes back?** *Answer: Basket*



First-Order, True-Belief Question

Sally-Anne Test

(Wimmer and Perner, 1983;
Baron Cohen et al., 1985;
Le, Boureau & Nickel 2019 for LLMs)

*Alice and Bob are in a room, where
there is a basket and a box.*

*Alice puts an apple in the basket,
and then leaves the room.*

Bob then moves the apple to the box.

**1. Where will Bob search for the
apple? Answer: Box**



First-Order, True-Belief Question

**2. Where does Bob think that Alice
will search for the apple when she
comes back? Answer: Basket**



Second-Order, False-Belief Question

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

Theory of Mind-specific domain language: example

<story start>

Anne entered the kitchen.

Theory of Mind-specific domain language: example

<story start>

Anne entered the kitchen.

**update
#1**

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



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.

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



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.

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
...

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.

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**

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.

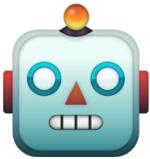
Charles entered the kitchen.

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^**

<story start>

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

<story start>

Anne entered the kitchen.
Beth entered the kitchen.

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

<story start>

Anne entered the kitchen.
Beth entered the kitchen.

update
#1

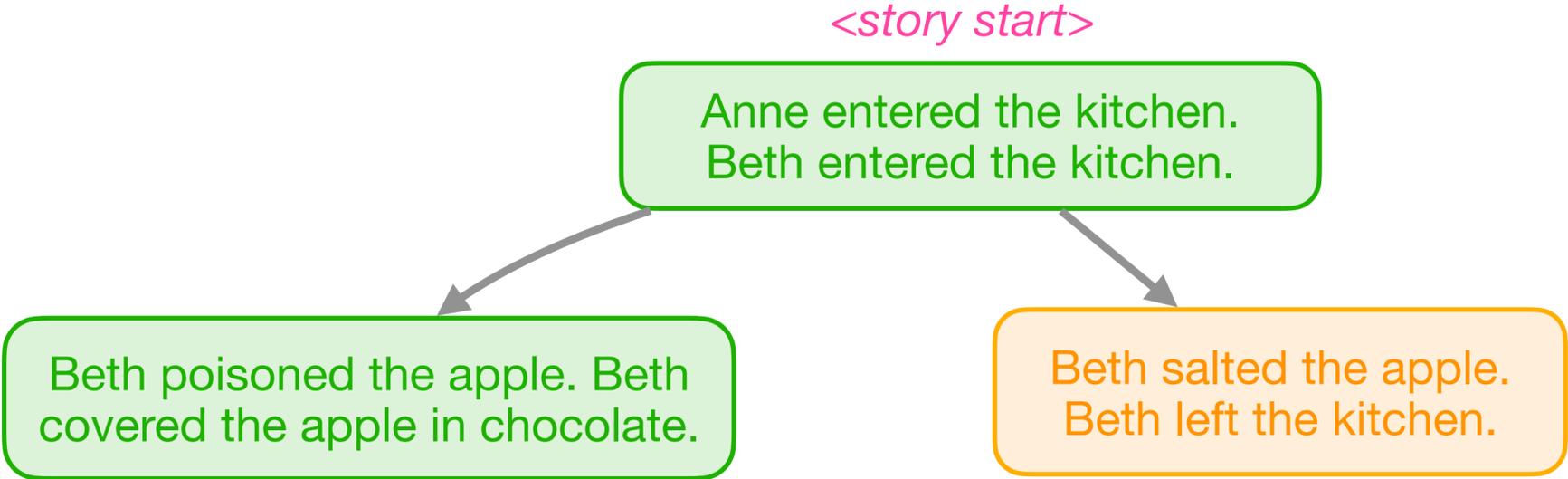
worldState[Beth, location] -> kitchen 
belief[Anne, Beth, location] -> kitchen
...

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

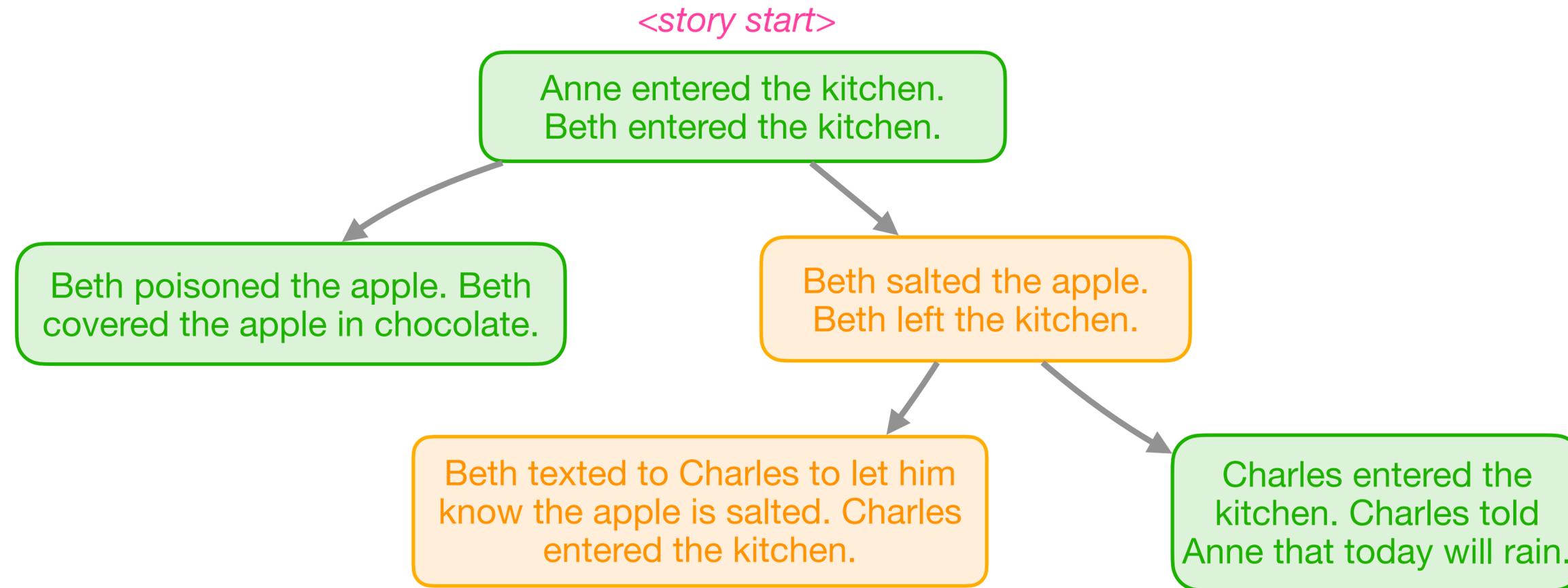
<story start>

Anne entered the kitchen.
Beth entered the kitchen.

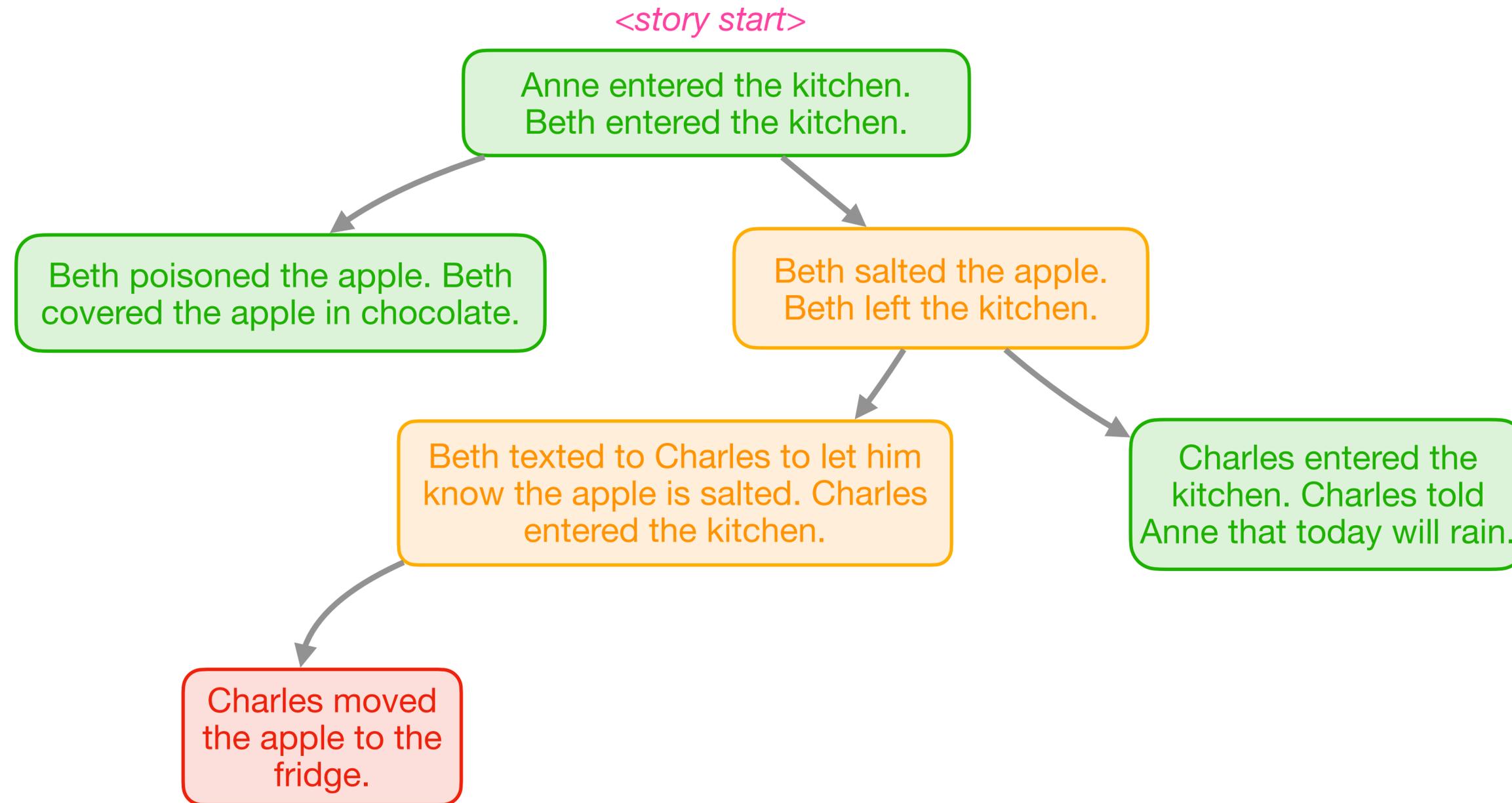
Adversarial story generation: searching for difficult stories with A^*



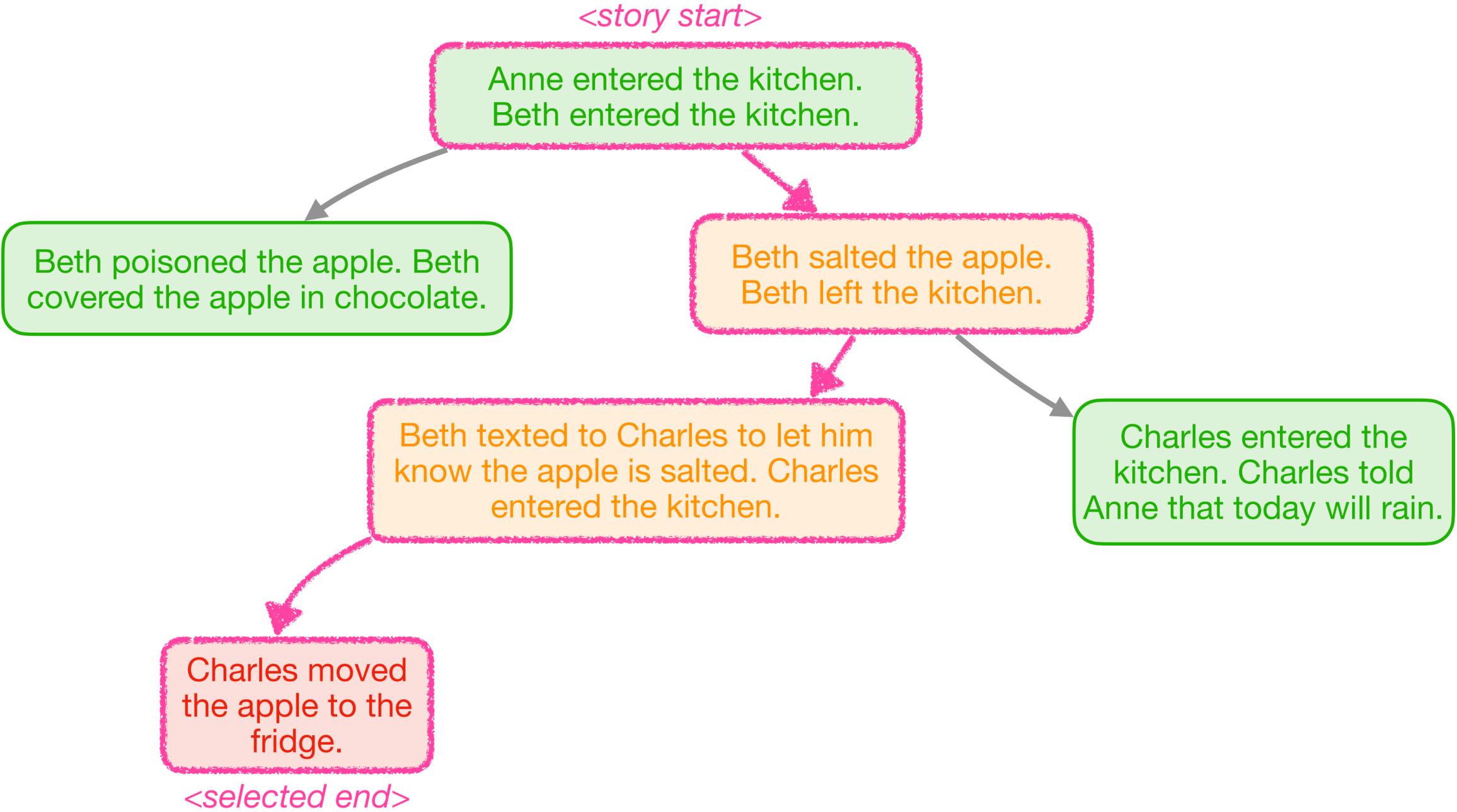
Adversarial story generation: searching for difficult stories with A^*



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

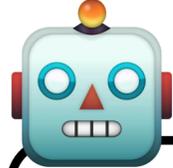


Adversarial story generation: searching for difficult stories with A^*



TrackTheMind: *full setup*

TrackTheMind: *full setup*



A. *Sample story context*

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.

Location: Restaurant kitchen.

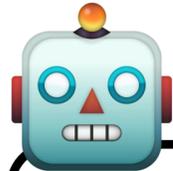
Alternative location: Walk-in pantry.

Object: apple.

Plausible containers: wooden crate; fridge.

Discussion topics: food safety protocols; menu changes.

TrackTheMind: *full setup*



A. *Sample story context*

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.

Location: Restaurant kitchen.

Alternative location: Walk-in pantry.

Object: apple.

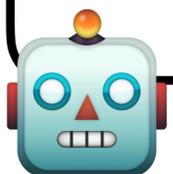
Plausible containers: wooden crate; fridge.

Discussion topics: food safety protocols; menu changes.



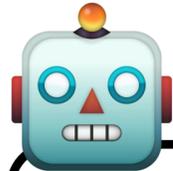
Visible state changes: covering the apple in chocolate; peeling the apple, ...

Invisible state changes: salting the apple; poisoning the apple; ...



B. *Sample state updates*

TrackTheMind: *full setup*



A. Sample story context

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.

Location: Restaurant kitchen.

Alternative location: Walk-in pantry.

Object: apple.

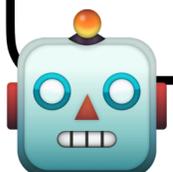
Plausible containers: wooden crate; fridge.

Discussion topics: food safety protocols; menu changes.



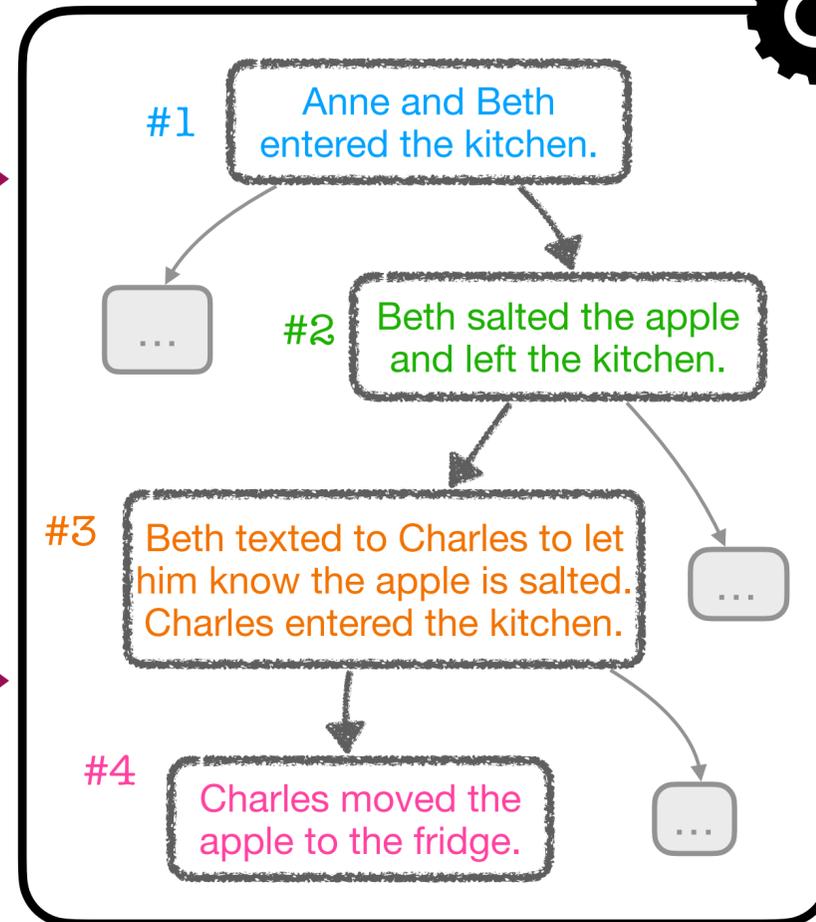
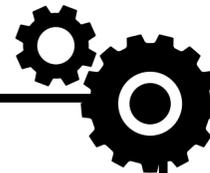
Visible state changes: covering the apple in chocolate; peeling the apple, ...

Invisible state changes: salting the apple; poisoning the apple; ...



B. Sample state updates

C. Search for difficult story structures with our mental-state tracker



TrackTheMind: *full setup*

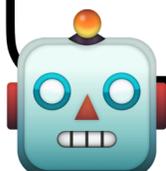


A. Sample story context

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.
Location: Restaurant kitchen.
Alternative location: Walk-in pantry.
Object: apple.
Plausible containers: wooden crate; fridge.
Discussion topics: food safety protocols; menu changes.

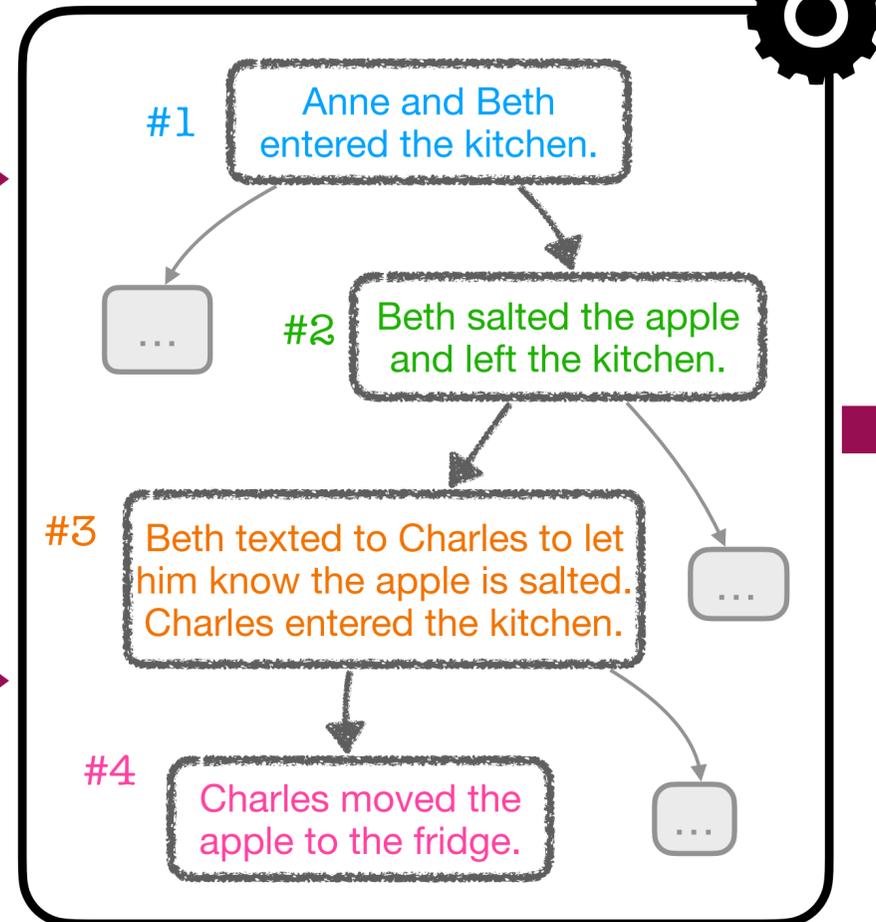


Visible state changes: covering the apple in chocolate; peeling the apple, ...
Invisible state changes: salting the apple; poisoning the apple; ...



B. Sample state updates

C. Search for difficult story structures with our mental-state tracker



D. Infill story incrementally

Context
Node #1
Node #2
#3

In the bustling kitchen of a high-end restaurant, the scent of freshly baked bread and simmering sauces filled the air, mingling with the hum of appliances and the soft clinking of pots and pans. *As the swinging kitchen doors parted, Anne strode in, her sharp eyes scanning the room to ensure every station was in full swing, and was closely followed by Beth, who made a beeline for the counter where a lone apple waited to be transformed into the evening's dessert masterpiece. Beth's skilled hands moved with precision, sprinkling a pinch of salt onto the apple's tender flesh to draw out its natural sweetness. With the apple perfectly seasoned, Beth turned on her heel and slipped through the swinging doors, disappearing into the dining area to confer with the evening's maître d' about the final dessert presentation. Beth quickly pulled her phone from her pocket and shot off a text to Charles - "Apple's salted".*

TrackTheMind: *full setup*



A. Sample story context

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.
Location: Restaurant kitchen.
Alternative location: Walk-in pantry.
Object: apple.
Plausible containers: wooden crate; fridge.
Discussion topics: food safety protocols; menu changes.

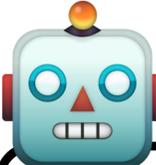
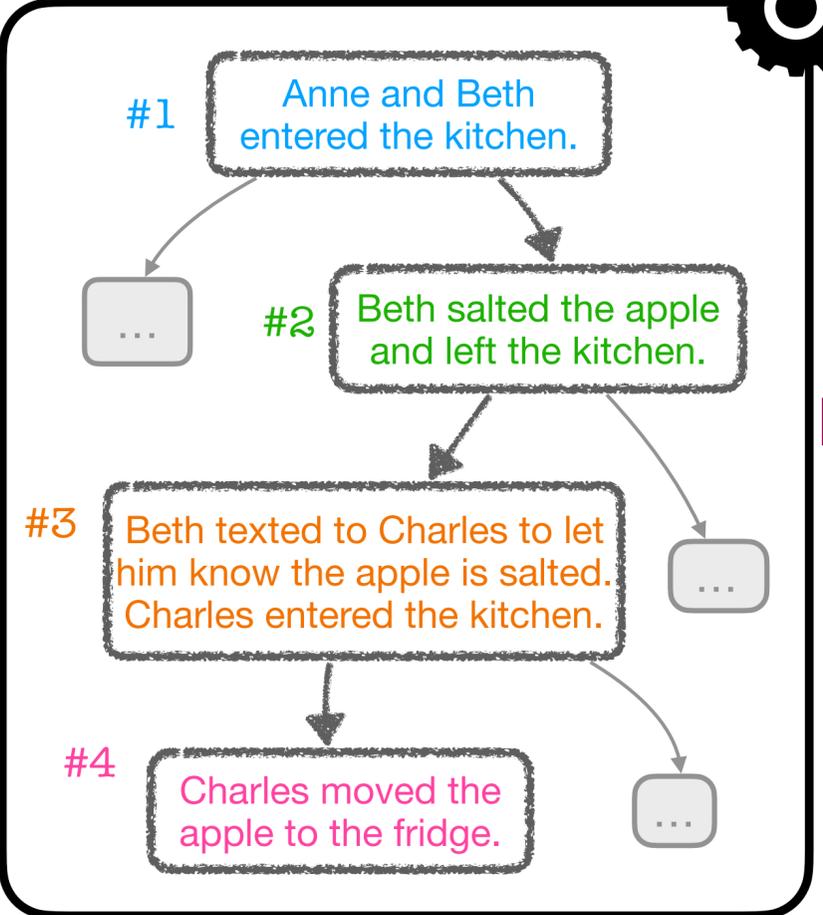


Visible state changes: covering the apple in chocolate; peeling the apple, ...
Invisible state changes: salting the apple; poisoning the apple; ...



B. Sample state updates

C. Search for difficult story structures with our mental-state tracker



D. Infill story incrementally

Context
Node #1
Node #2
#3

In the bustling kitchen of a high-end restaurant, the scent of freshly baked bread and simmering sauces filled the air, mingling with the hum of appliances and the soft clinking of pots and pans. *As the swinging kitchen doors parted, Anne strode in, her sharp eyes scanning the room to ensure every station was in full swing, and was closely followed by Beth, who made a beeline for the counter where a lone apple waited to be transformed into the evening's dessert masterpiece. Beth's skilled hands moved with precision, sprinkling a pinch of salt onto the apple's tender flesh to draw out its natural sweetness. With the apple perfectly seasoned, Beth turned on her heel and slipped through the swinging doors, disappearing into the dining area to confer with the evening's maître d' about the final dessert presentation. Beth quickly pulled her phone from her pocket and shot off a text to Charles - "Apple's salted".*

TrackTheMind: *full setup*



A. Sample story context

People: Anne, a head chef; Beth, a pastry chef; Charles, a line cook.
Location: Restaurant kitchen.
Alternative location: Walk-in pantry.
Object: apple.
Plausible containers: wooden crate; fridge.
Discussion topics: food safety protocols; menu changes.

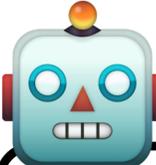
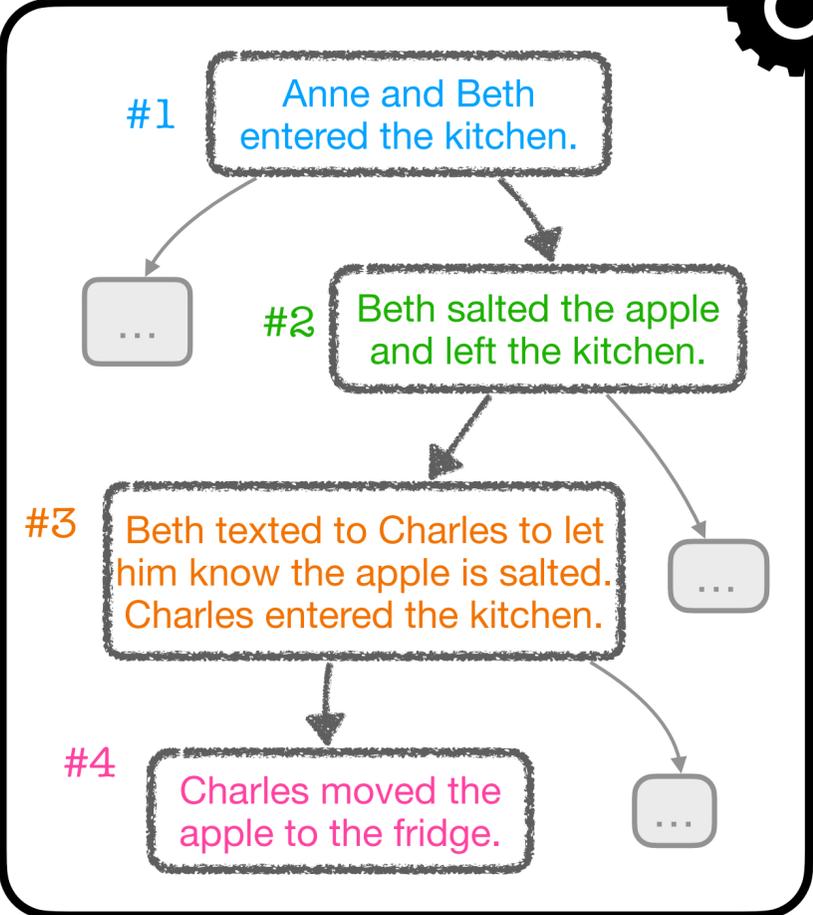


Visible state changes: covering the apple in chocolate; peeling the apple, ...
Invisible state changes: salting the apple; poisoning the apple; ...



B. Sample state updates

C. Search for difficult story structures with our mental-state tracker



D. Infill story incrementally

Context
Node #1
Node #2
#3

In the bustling kitchen of a high-end restaurant, the scent of freshly baked bread and simmering sauces filled the air, mingling with the hum of appliances and the soft clinking of pots and pans. *As the swinging kitchen doors parted, Anne strode in, her sharp eyes scanning the room to ensure every station was in full swing, and was closely followed by Beth, who made a beeline for the counter where a lone apple waited to be transformed into the evening's dessert masterpiece. Beth's skilled hands moved with precision, sprinkling a pinch of salt onto the apple's tender flesh to draw out its natural sweetness. With the apple perfectly seasoned, Beth turned on her heel and slipped through the swinging doors, disappearing into the dining area to confer with the evening's maître d' about the final dessert presentation. Beth quickly pulled her phone from her pocket and shot off a text to Charles - "Apple's salted".*

[A subset of] TrackTheMind results

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

Improving Reasoning

At Training time vs. Inference time

ExploreToM: 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)

ExploreToM: 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?

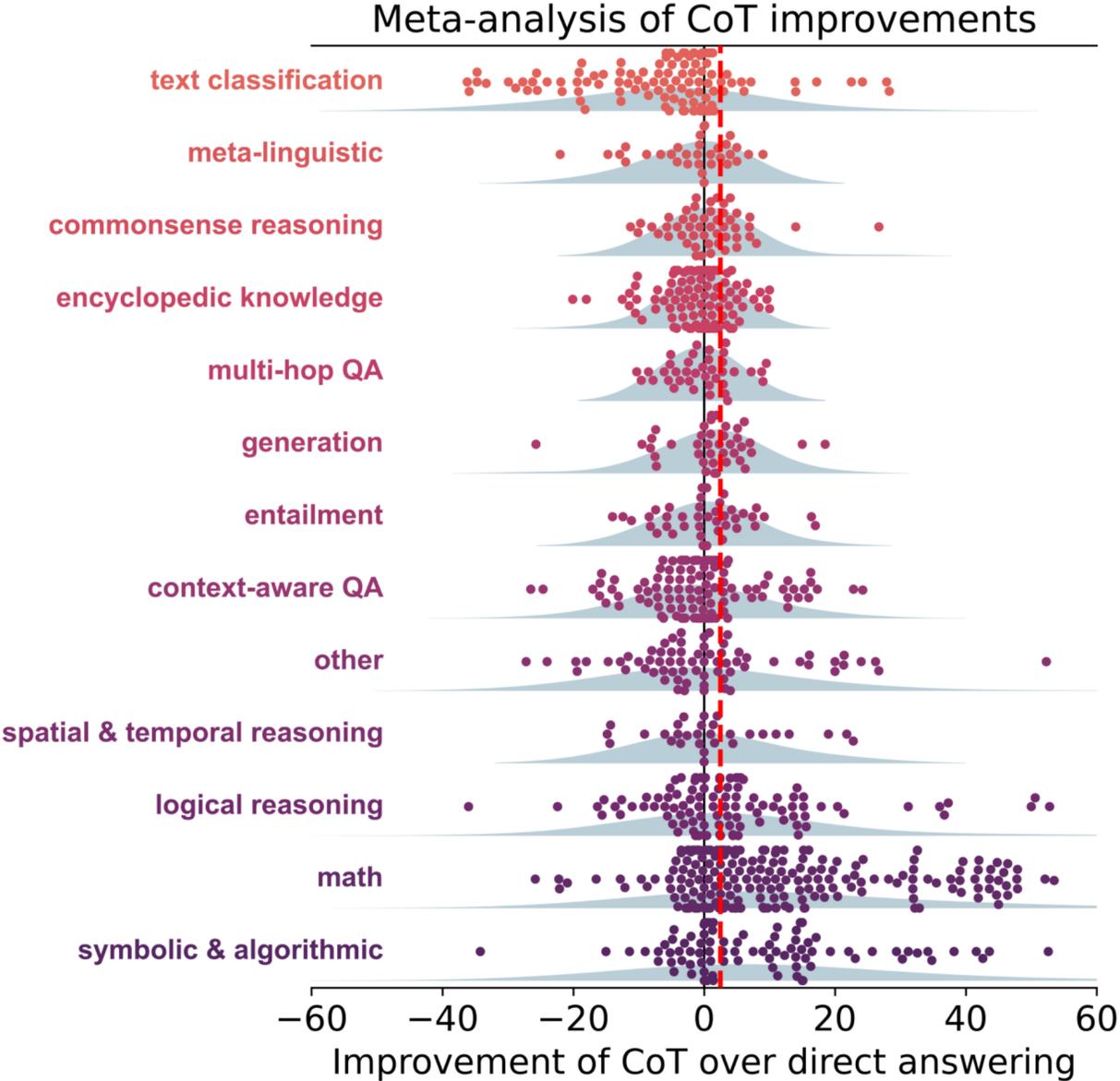
	ToMi	Hi-ToM	BigToM	OpenToM (F1)	FANToM
Llama-3.1 8B Instruct	68%	30%	75%	.39	0.3%
TRACKTHEMIND-8B	94% (+26)	52% (+22)	79% (+4)	.42 (+.03)	0.5% (+0.02)

Example of a possible action plan when even frontier models cannot generate good data for knowledge distillation!

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*



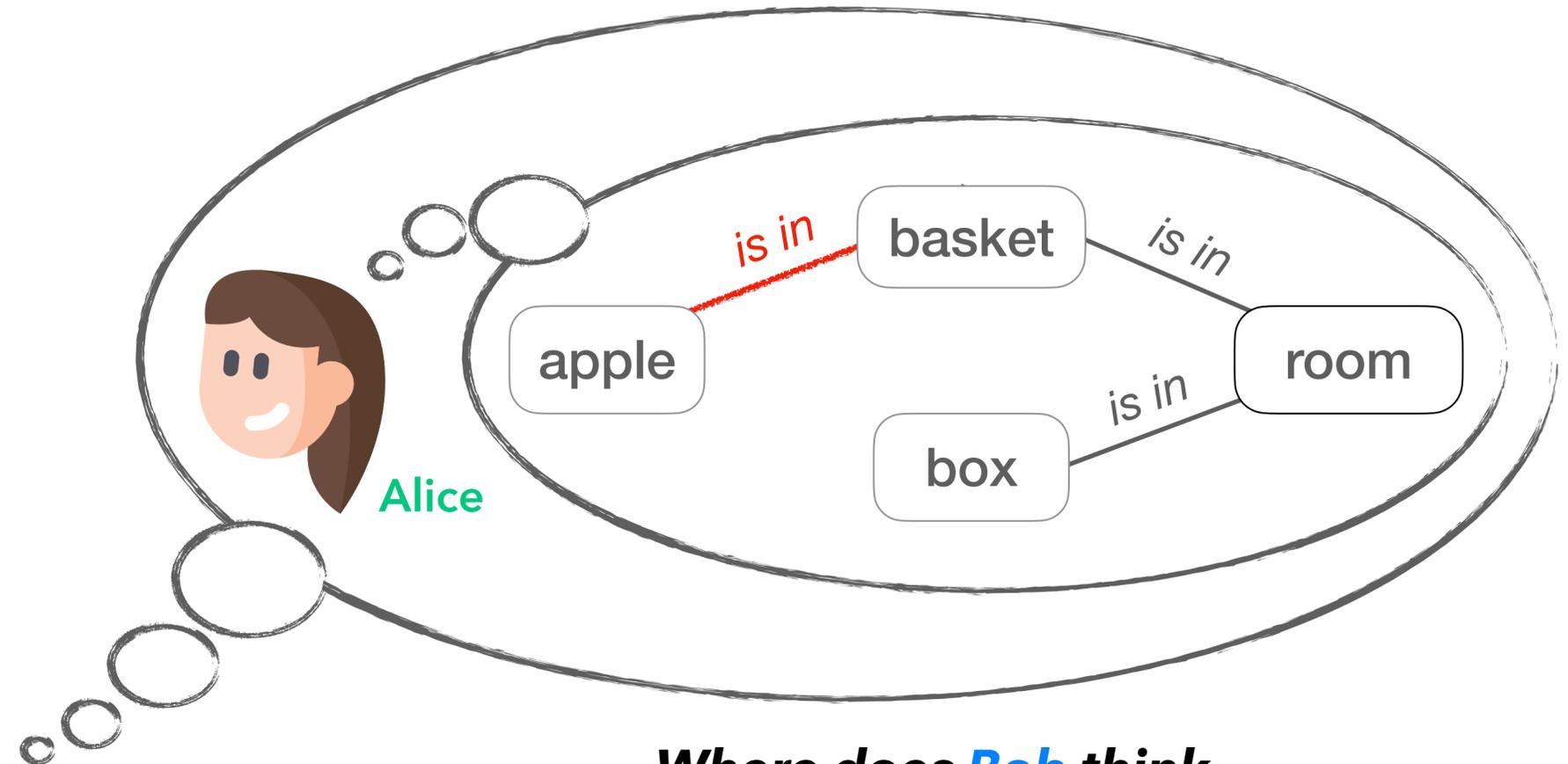
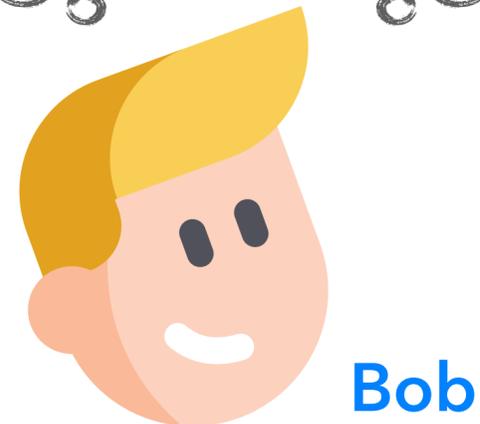
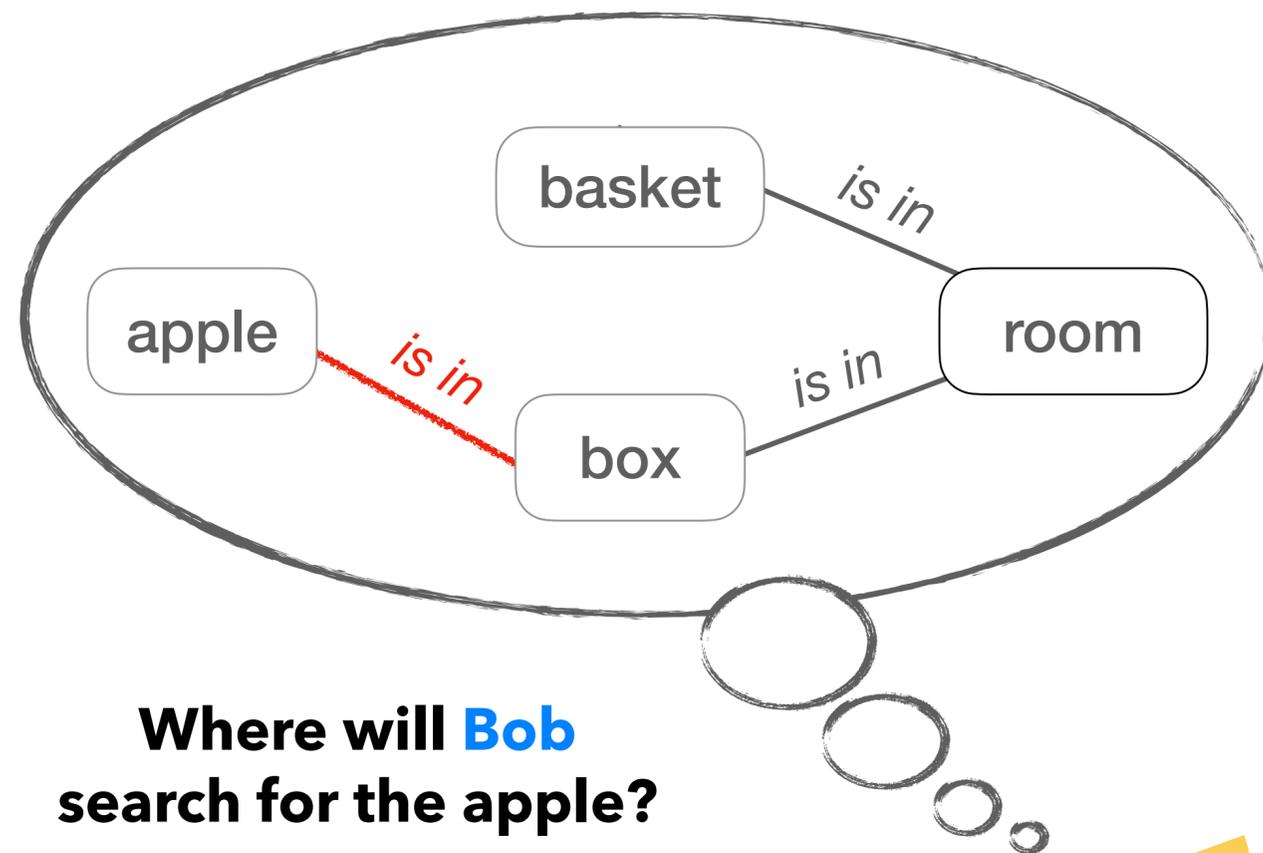
*Yejin
Choi*



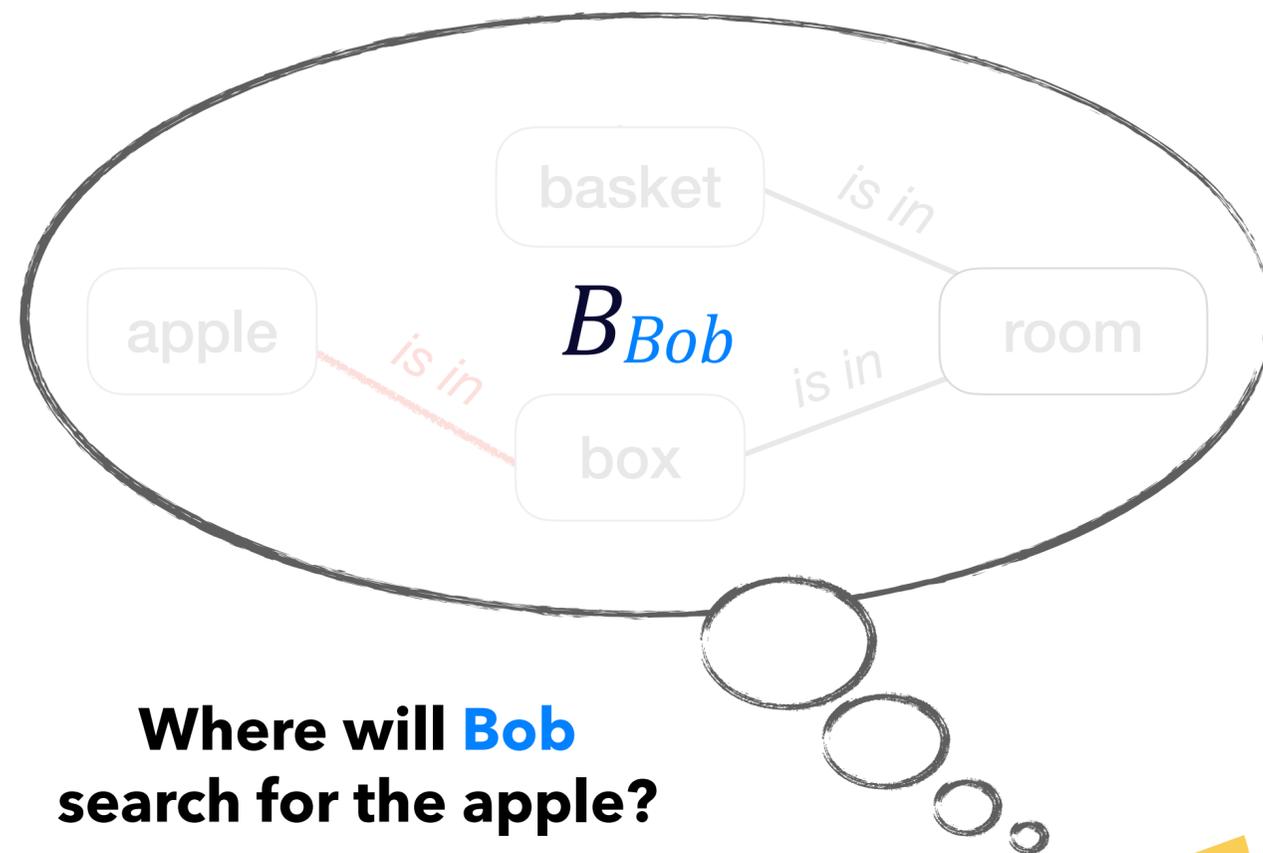
*Yulia
Tsvetkov*



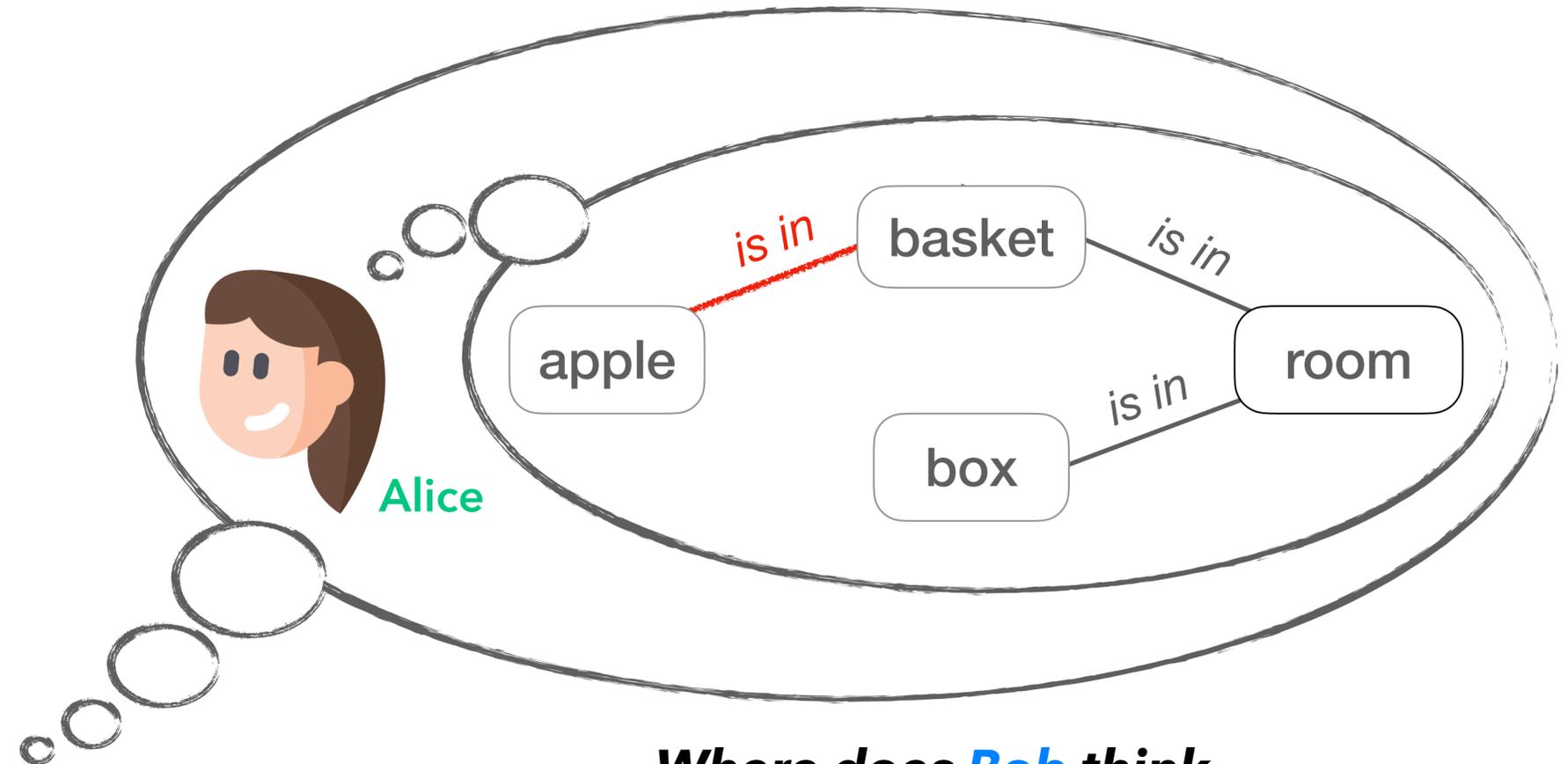
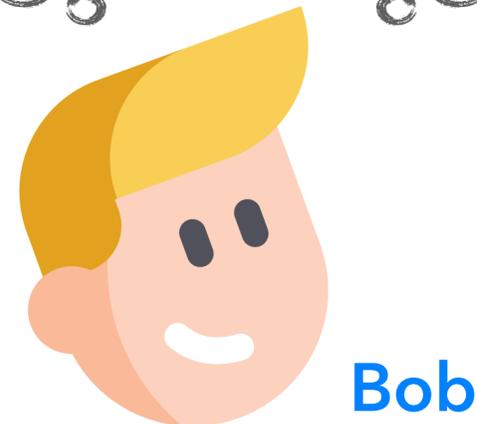
Graphical Representations of Local Context



Graphical Representations of Local Context

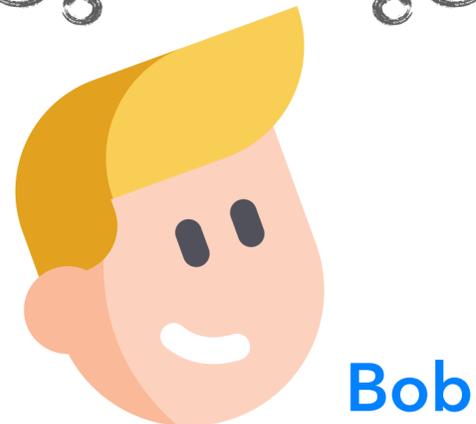
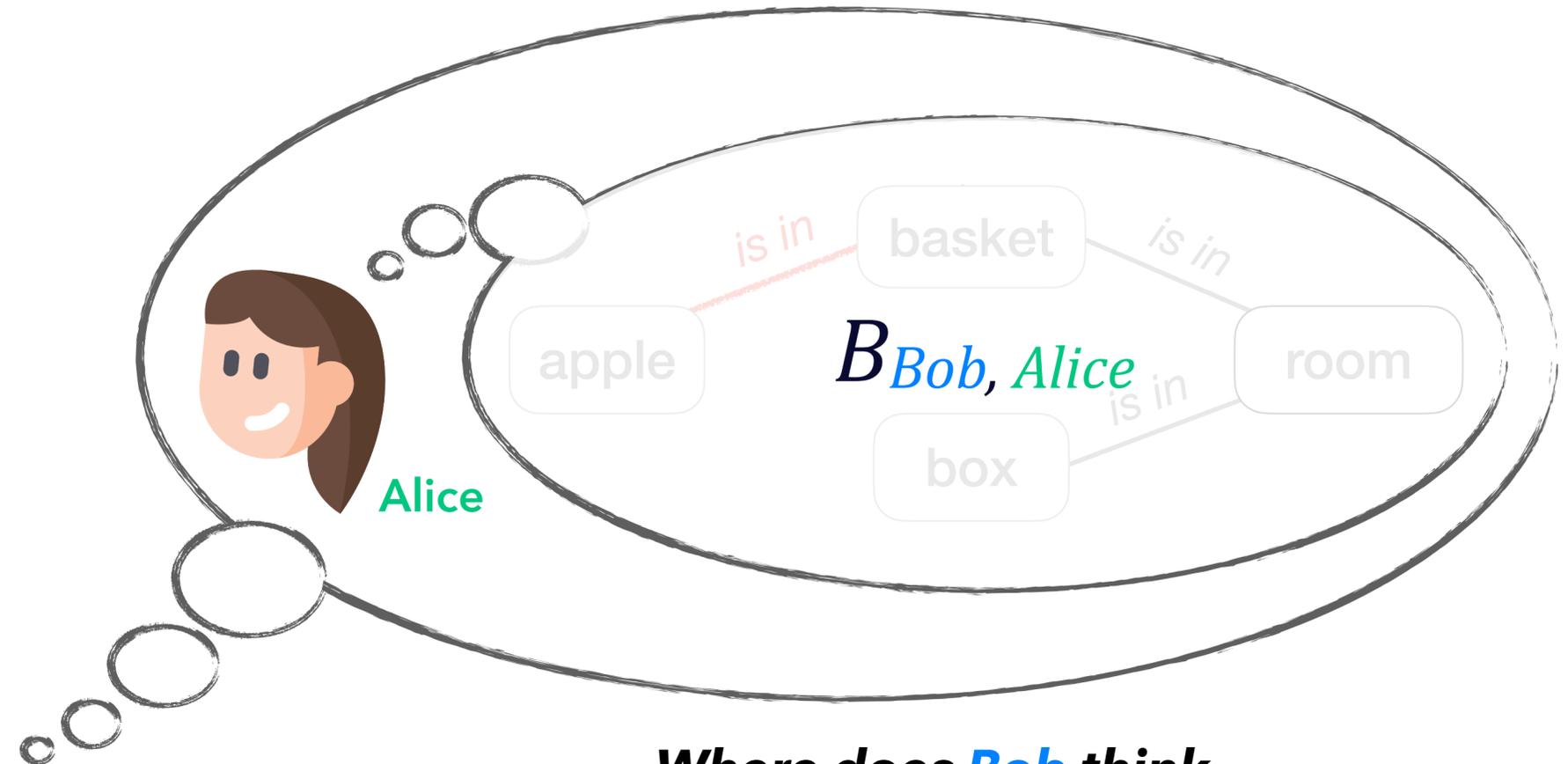
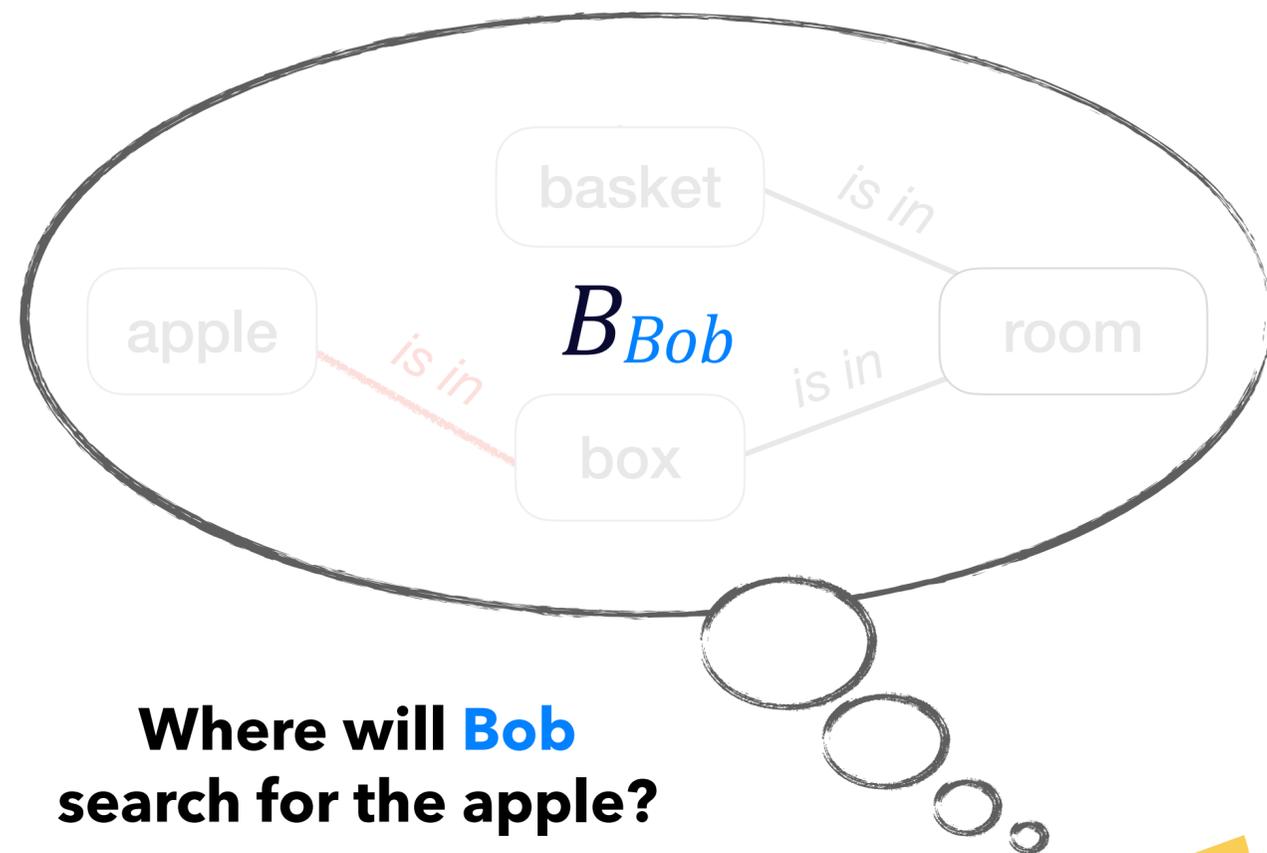


Where will **Bob** search for the apple?



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

Graphical Representations of Local Context



Symbolic ToM Overview

*Where does
Alice think
that **Bob**
will search for
the apple?*

Symbolic ToM Overview

*Where does
Alice think
that **Bob**
will search for
the apple?*

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

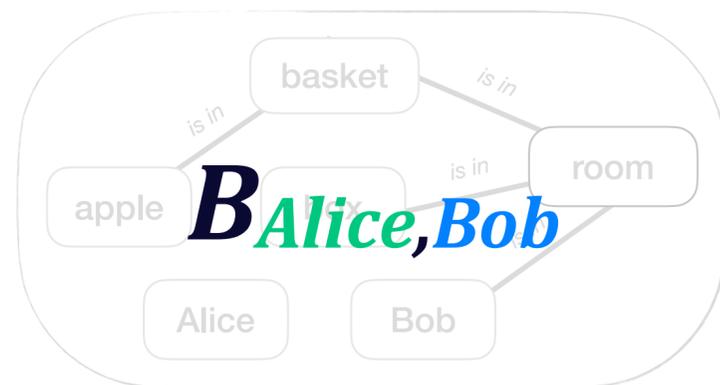
Symbolic ToM Overview

**Where does
Alice think
that Bob
will search for
the apple?**

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

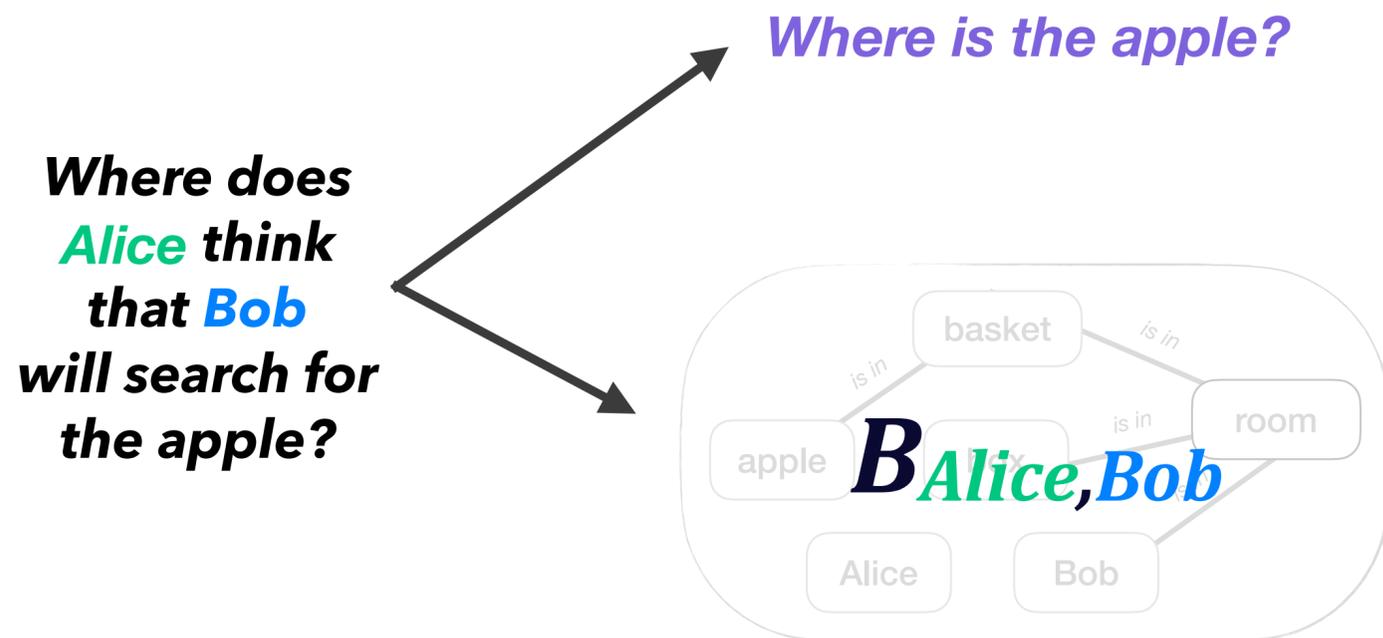
Symbolic ToM Overview

Where does
Alice think
that *Bob*
will search for
the apple?



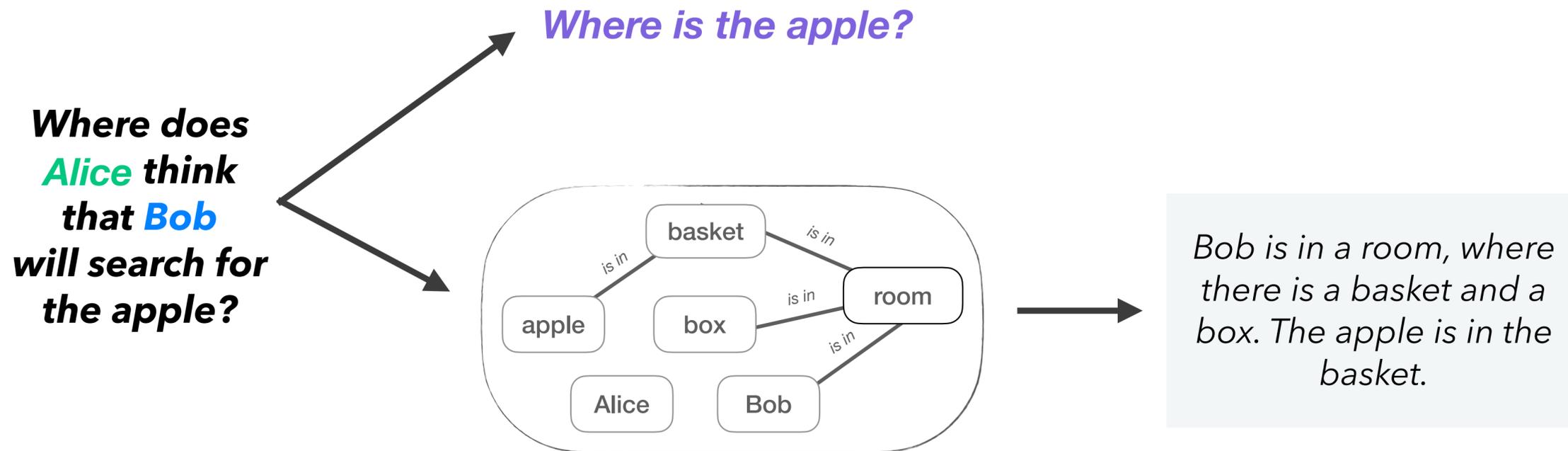
1. Detect entities in question, retrieve belief graph and perform recursion over the question

Symbolic ToM Overview



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

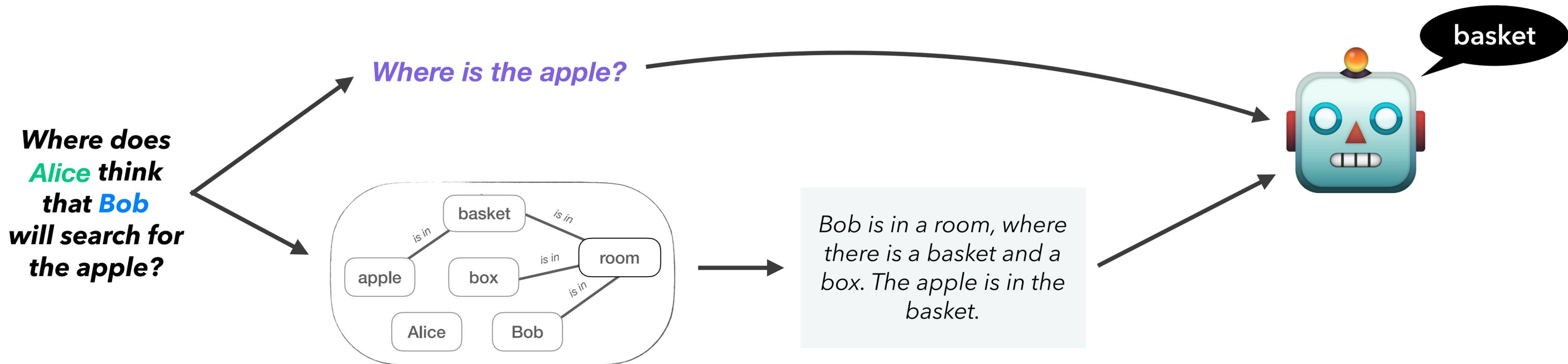
Symbolic ToM Overview



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

2. Retrieve sentences captured by the graph

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
<i>SYMBOLICTOM + Off-the-shelf models</i>			
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)
<i>Supervised models</i>			
TTT	49	65	78
Finetuned GPT3	51	68	32

Results: Out-of-Domain Performance

Story Structure
Generalization

	D_1	D_2	D_3
<i>SYMBOLICTOM + Off-the-shelf models</i>			
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)
<i>Supervised models</i>			
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