Learning to Reason Deductively: Towards Interpretable Math Word Problem Solving

Allan Jie

Joint work with Jierui Li and Prof. Wei Lu

ByteDance AI Lab

TEXAS
The University of Texas at Austin

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Accepted by ACL-2022

Code: https://github.com/allanj/Deductive-MWP
Paper: https://arxiv.org/abs/2203.10316

Motivation: Reasoning

1. Pre-trained language models achieve promising performance in many NLP tasks.

2. But still suffers in multi-step reasoning tasks

   (a) Math word problem solving is a straightforward application to measure machine learning models' ability in understanding language.

TLDR: Chain-of-Thoughts Prompts (Few-shot)

Standard Prompting	Chain of thought prompting
Example Input	**Example Input**
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?	Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
Example Output	**Example Output**
A: The answer is 11.	Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.
Prompt	**Prompt**
The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?	The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
Model Response	**Model Response**
The answer is 50.	The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20 = 3. They bought 6 more apples, so they have 3+6=9. The answer is 9.

PaLM: 540-Billion Parameter Language Model with Pathways (Google)

Problem Description

Question (Input): *In a division sum , the remainder is 8 and the divisor is 6 times the quotient and is obtained by adding 3 to the thrice of the remainder. What is the dividend?*

Mathematical Expression (Output): ((8 x 3 + 3) x (8 x 3 + 3)/6) + 8

Answer: 129.5

Assume positions of quantities are known, and only consider "+", "-", "*", "/", "^"

Sequence-to-Sequence Models

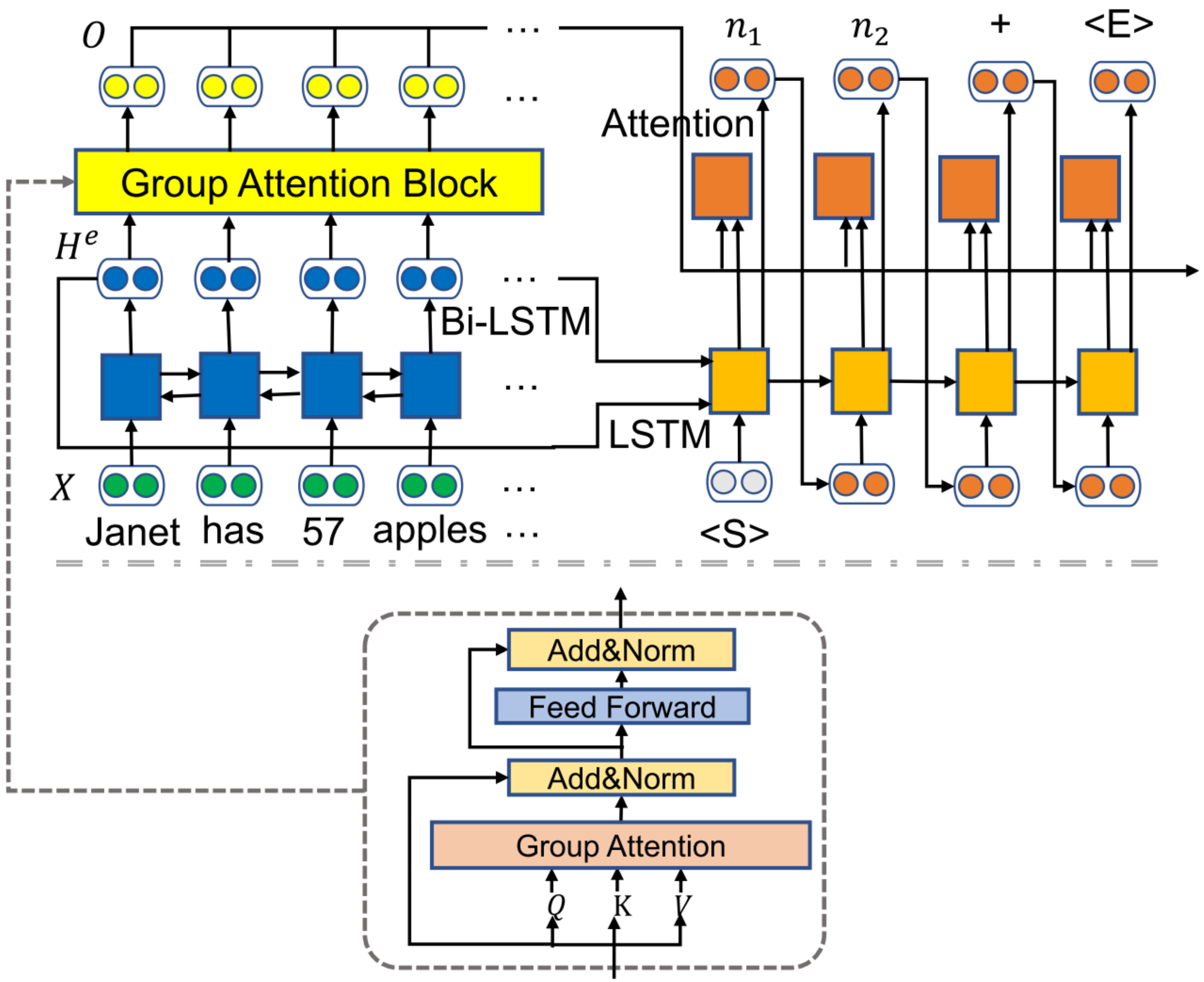


Figure taken from Li et al., (2019) ACL. "*Modeling Intra-Relation in Math Word Problems with Different Functional Multi-Head Attentions*"

Pros ✅: easy to implement and general for different types of problems

Cons ❎:

1. Performance is far from satisfactory
2. Lack of interpretability for prediction.

Note: this direction is still popular because of Transformers (Shen et al., 2021)

Tree-based Models

Question: *In a division sum , the remainder is 8 and the divisor is 6 times the quotient and is obtained by adding 3 to the thrice of the remainder. What is the dividend?*

Answer: 129.5

Mathematical Expression:

((8 x 3 + 3) x (8 x 3 + 3)/6) + 8

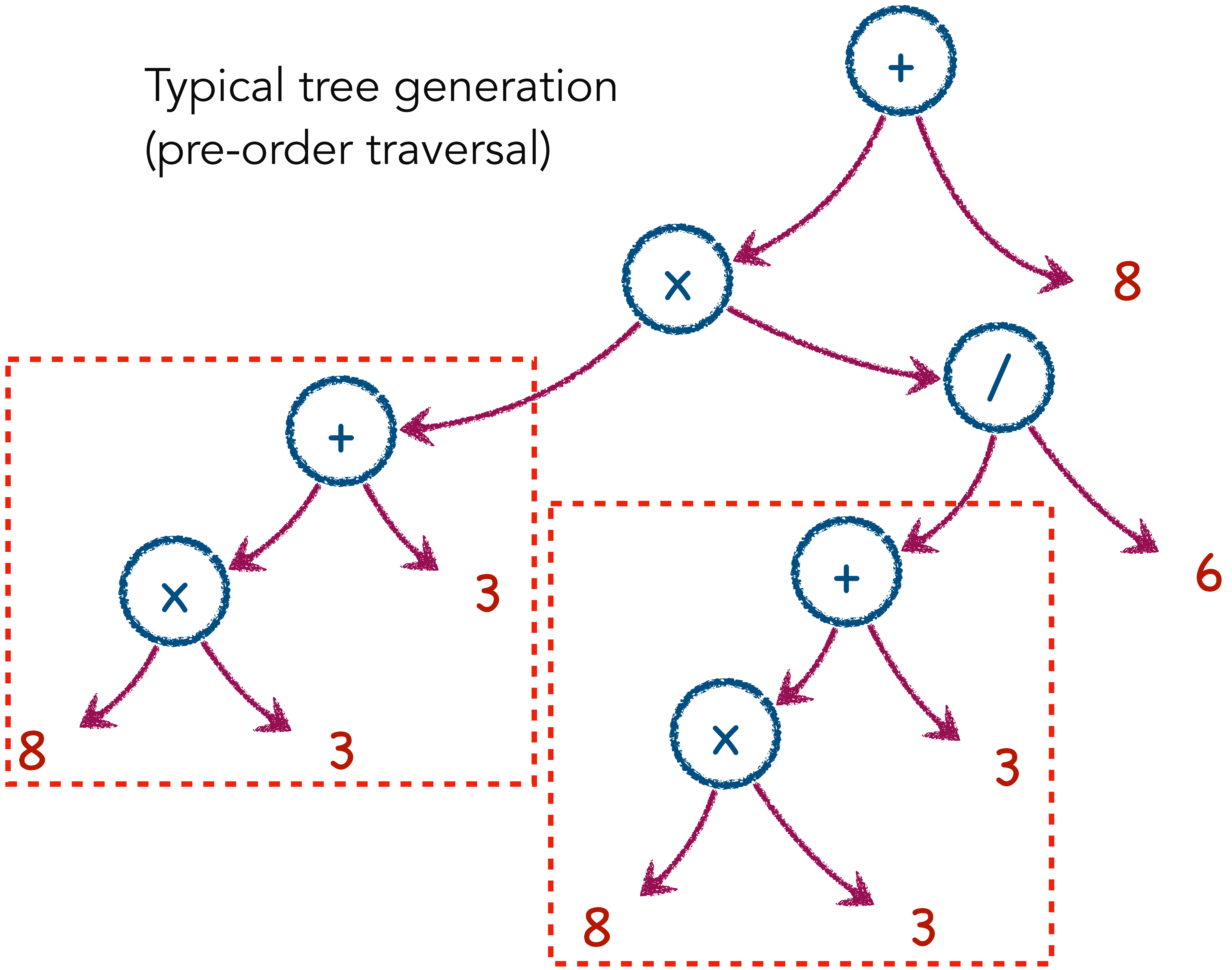
Pros ✅: generate tree structures

Cons ❎:

1. Generation process is still counter-intuitive

2. Repetitive computation

Typical tree generation
(pre-order traversal)



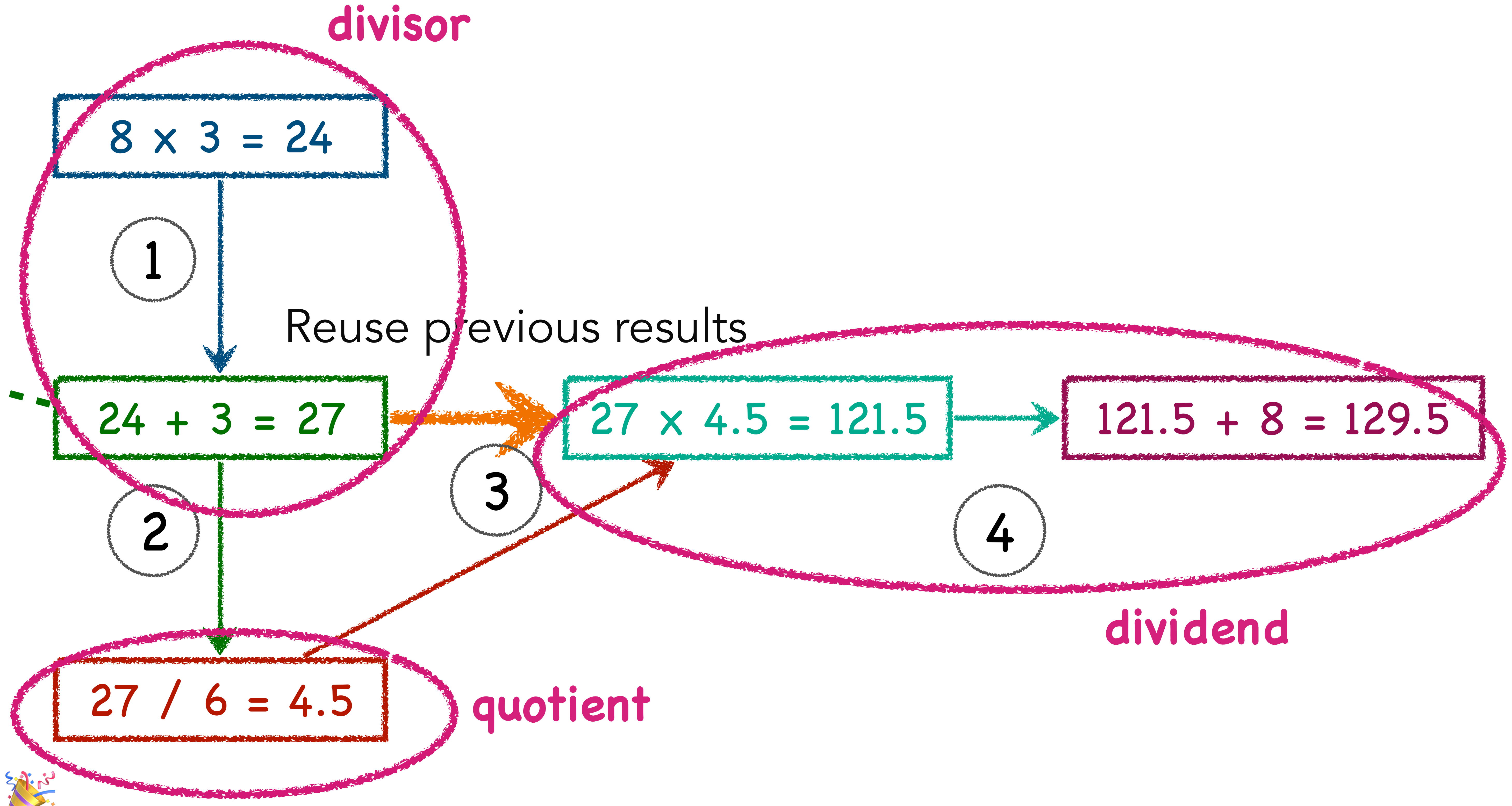
Deductive Systems for Math Word Problem Solving

Question: *In a division sum , the remainder is 8 and the divisor is 6 times the quotient and is obtained by adding 3 to the thrice of the remainder. What is the dividend?*

Answer: 129.5

Mathematical Expression:

((8 x 3 + 3) x (8 x 3 + 3)/6) + 8



1. Interpretable and less steps for computation 🎉
2. Deductive process, different from traditional tree-based generation
    a) Also allows us to inference from intermediate state (i.e., step)
3. Able to reuse previous calculated results
4. Generate the expression directly, rather than single token

Method: Deductive Reasoner

1. Input: quantities $Q = Q^{(t=0)} = q_1, q_2, \cdots., q_m$ (including constants)

2. $e^{(t)}_{i,j,op} = q_i \xrightarrow{op} q_j \quad q_i, q_j \in \mathcal{Q}^{(t-1)}$ op is the operator(e.g., "+", "-")

$$\textbf{input:} \quad q \text{ in } \mathcal{Q}^{(0)}$$

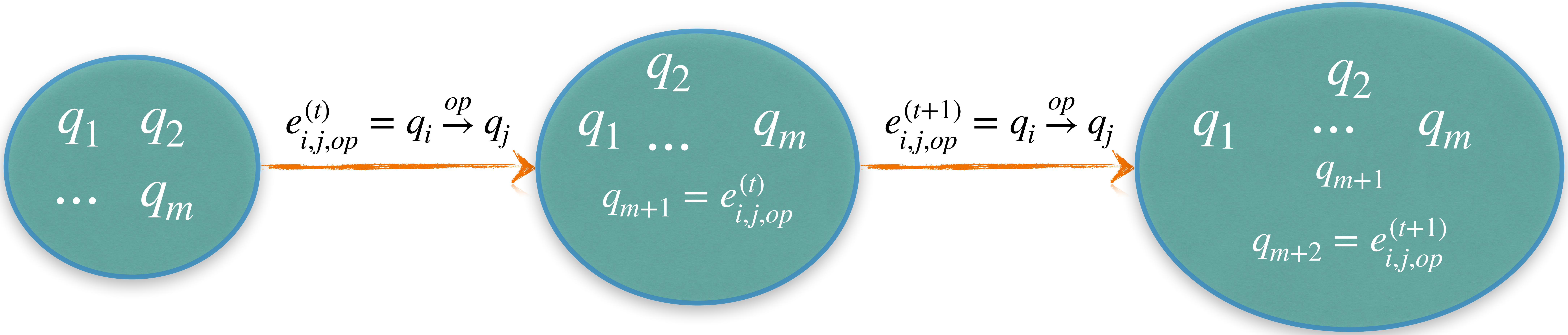
$$\textbf{axiom:} \quad 0 : \langle q_1, \cdots, q_{|\mathcal{Q}^{(0)}|} \rangle$$

$$q_i \xrightarrow{op} q_j : \frac{t : \langle q_1, \cdots, q_{|\mathcal{Q}^{(t-1)}|} \rangle}{t+1 : \langle q_1, \cdots, q_{|\mathcal{Q}^{(t-1)}|} \mid q_{|\mathcal{Q}^{(t)}|} := e^{(t)}_{i,j,op} \rangle}$$

Method: Deductive Reasoner

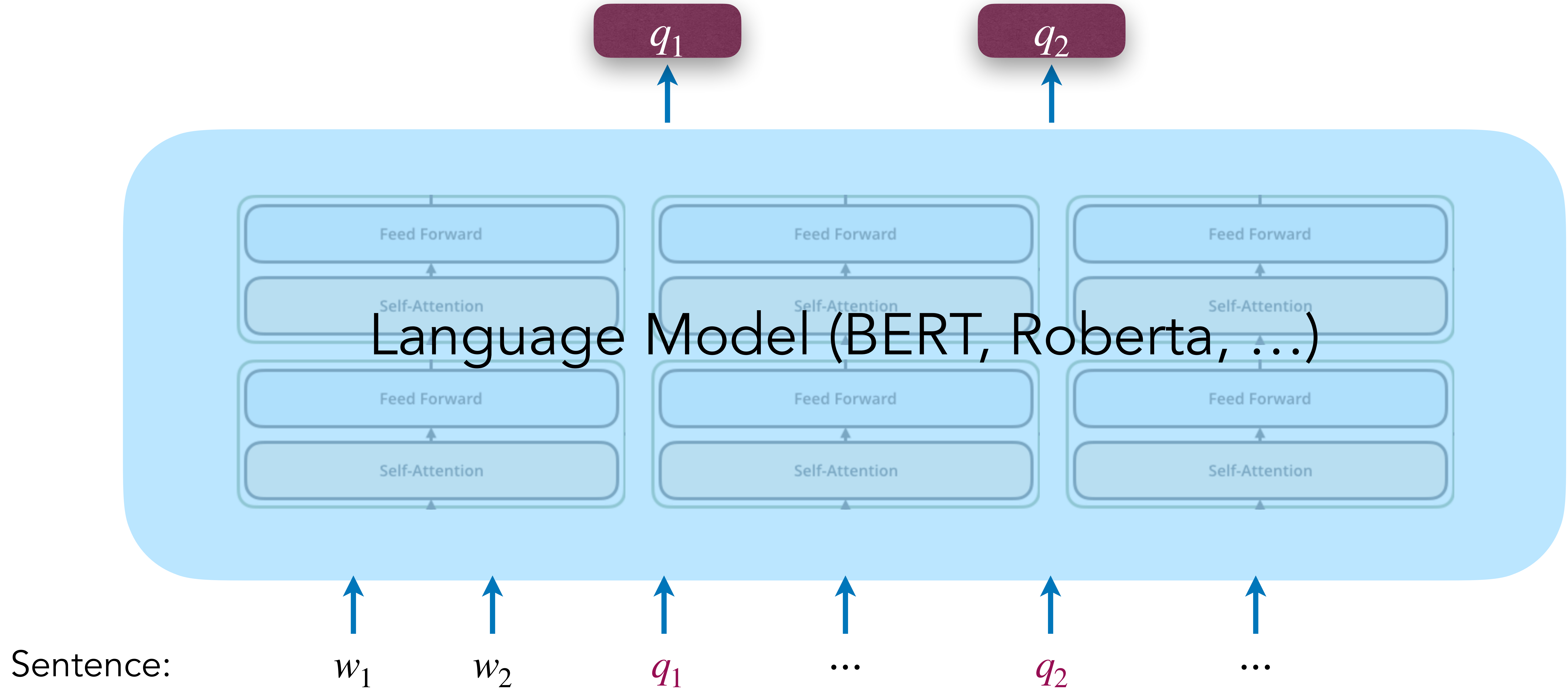
1. Input: quantities $Q = Q^{(t=0)} = q_1, q_2, \cdots., q_m$ (including constants)

2. $e^{(t)}_{i,j,op} = q_i \overset{op}{\rightarrow} q_j$ $\quad q_i, q_j \in \mathcal{Q}^{(t-1)}$ op is the operator(e.g., "+", "-")



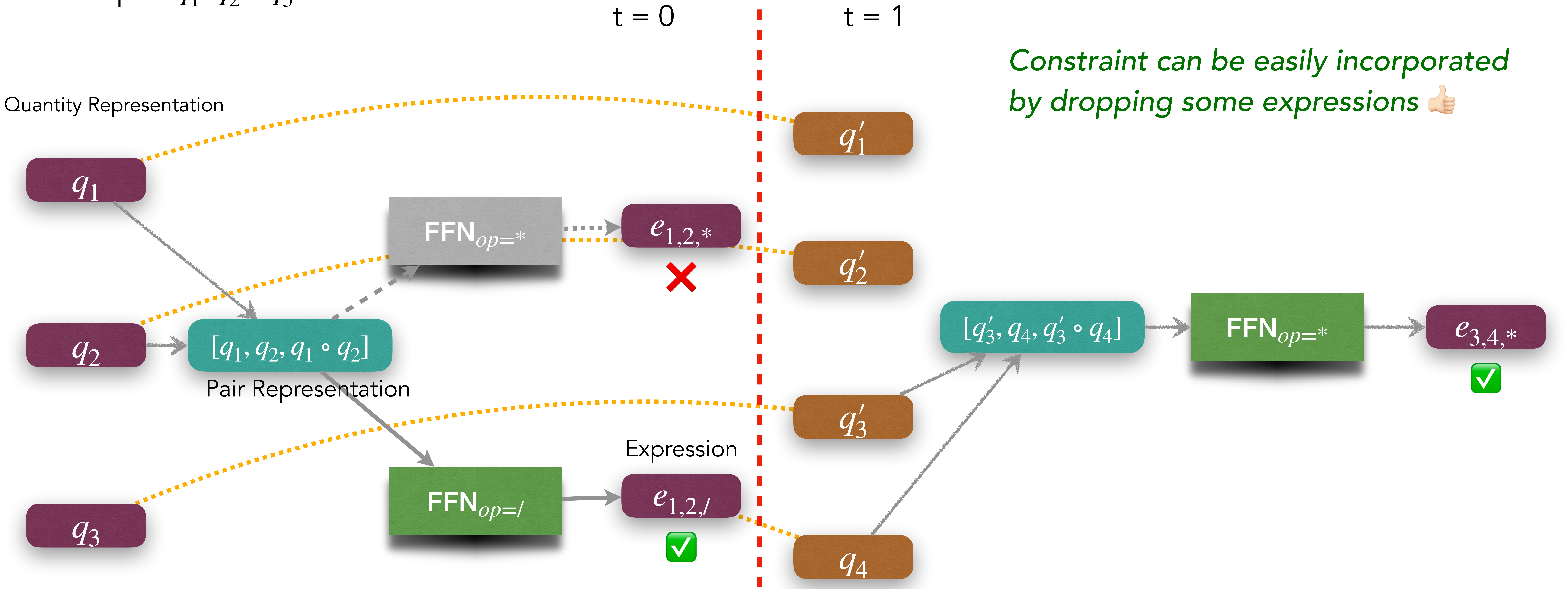
Model Architecture (Text Encoder)

Obtain quantity representation



Model Architecture (Inference)

Example: $q_1/q_2 * q_3$

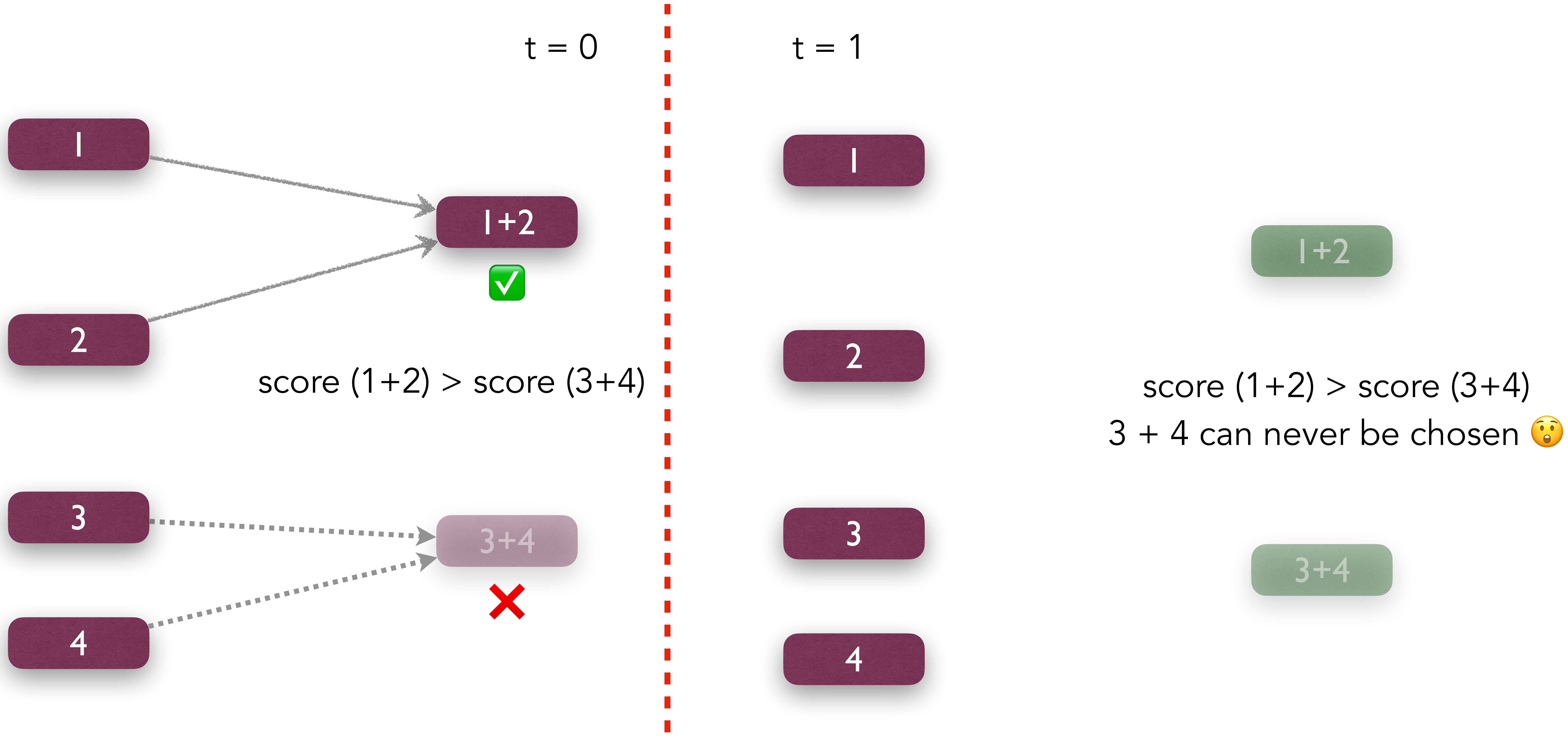


t = 0, #possible expression = $\binom{3}{2} \times |op|$

t = 0, #possible expression = $\binom{4}{2} \times |op|$

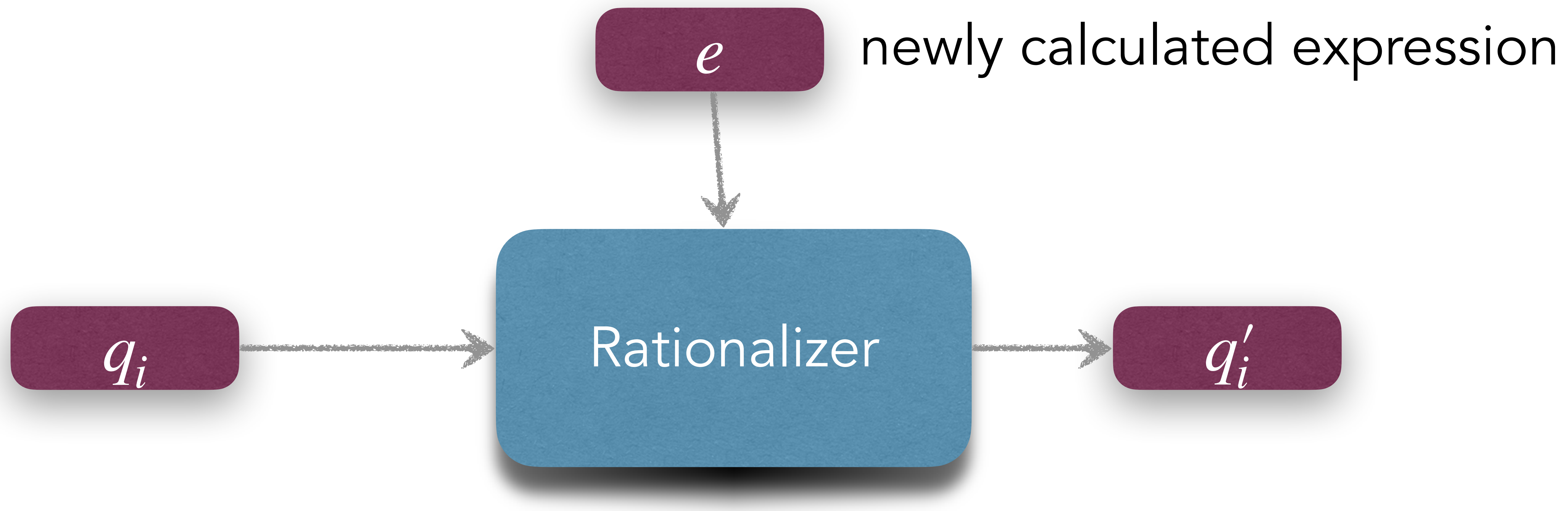
Why q_1' (should) be different from q_1?

1. Consider a simple example: (1 + 2) * (3 + 4) * (5 + 6)



Rationalizer

1. Rationalizing the quantity representation



Rationalizer	Mechanism
Self-Attention	Attention(Q = [qi , e] , K = [qi , e] ,V = [qi , e])
GRU cell	GRU_Cell(input = qi , previous hidden = e)

Likely to return the same representation

Training

1. Similar to training sequence-to-sequence model, accumulating loss at each time step.

2. The search space $\mathscr{H}$ at each step is different! 📝

    1. The space allows us to impose constraint (e.g., negative results are not allow, etc.)

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left(\max_{(i,j,op)\in\mathscr{H}^{(t)},\tau} \left[s_{\boldsymbol{\theta}}(e^{(t)}_{i,j,op}, \tau) \right] - s_{\boldsymbol{\theta}}(e^{(t)}_{i*,j*,op*}, \tau*) \right)$$

Experiments on Public Datasets

Dataset	**Amount**	**Description**	**Difficulties (from easiest 1 to 4 hardest)**
MAWPS	1987	English dataset developed in 2016, simple dataset for NLP to make early attempt on math word problem solving	1
Math23k	23162	Large-scale dataset in Math23k, mainly primary-school problems	2
MathQA	20121	Harder problems in different subjects, physics, science, etc.	3
SVAMP	4138	Carefully curated problems from MAWPS and ADV-DIVS to evaluate NLP models	4

Results


Previous Best
Our Deductive Reasoner
Value Accuracy
95
80
65
50
35
88.7
92.0
82.4
85.1
77.1
78.6
43.8
47.3
MAWPS
Math23K
MathQA
SVAMP

Fine-Grained Analysis on SVAMP dataset

1. We observe some intermediate results are negative on SVAMP. This dataset is hard because

   (a) Manually created to try to confuse the NLP models

   (b) Added extra irrelevant information (i.e., extra quantities)

   (c) Rephrasing some important statement (such as comparison)

2

```
{
   "question": "jake has 13 more apples and 17 fewer peaches than steven . steven has 8 peaches and 12 apples . how
many apples does jake have ?",
  "pred_equation": [
    ['8 - 17 = -9'], ['-9 + 13 =4']
  ],
  "gold_equation": [
    ['13 + 12 = 25']
  ]
}
```

Details on SVAMP dataset

Model	Value Accuracy	Description
Roberta-Graph2Tree	43.8	Previous best
BERT-Deductive Reasoner	35.3	Constraint: disallow the intermediate results have negative number
BERT-Deductive Reasoner + constraints	42.3	
Roberta-Deductive Reasoner	45.0	
Roberta-Deductive Reasoner + constraints	**47.3**	

What's really the difficulty?

We investigate the number of unused quantities, which can be regarded as irrelevant information that confuse the models. 🤔 🤔

	MAWPS	Math23k	MathQA	SVAMP
Samples (%) with Unused quantities	6.5	8.2	20.7	44.5
0 unused quantities	93.6	87.1	81.4	63.6
>= 1 unused quantities	-	62.1	67.4	27.0

Question Perturbation

Question: There are 255 apple trees in the orchard. Planting another 35 pear trees makes the number exactly the same as the apple trees. If every 20 pear trees are planted in a row, how many rows can be planted in total?

Answer: 11. **Gold Expression**: (255 - 35) / 20. **Predicted Expression**: (255 + 35) / 20

Deductive Scores:

Prob('255+35=260') = 0.068 > Prob('255-35=220') = 0.062

Question: There are 255 apple trees in the orchard. The number of pear trees are 35 fewer than the apple trees. If every 20 pear trees are planted in a row, how many rows can be planted in total?

Answer: 11. **Gold Expression**: (255 - 35) / 20. **Predicted Expression**: (255 - 35) / 20

Prob(255+35=260) = 0.061 < Prob(255-35=220) = 0.067

Takeaways

Pros

(1) Model is efficient and less number of steps required for inference

(2) Intuitive and the model makes interpretable predictions.

(3) Easily incorporate prior knowledge as constraints, which potentially can further boost the performance

(4) The underlying mechanism theoretically applies not only to MWP solving task but other tasks that potentially involves multi-step reasoning.

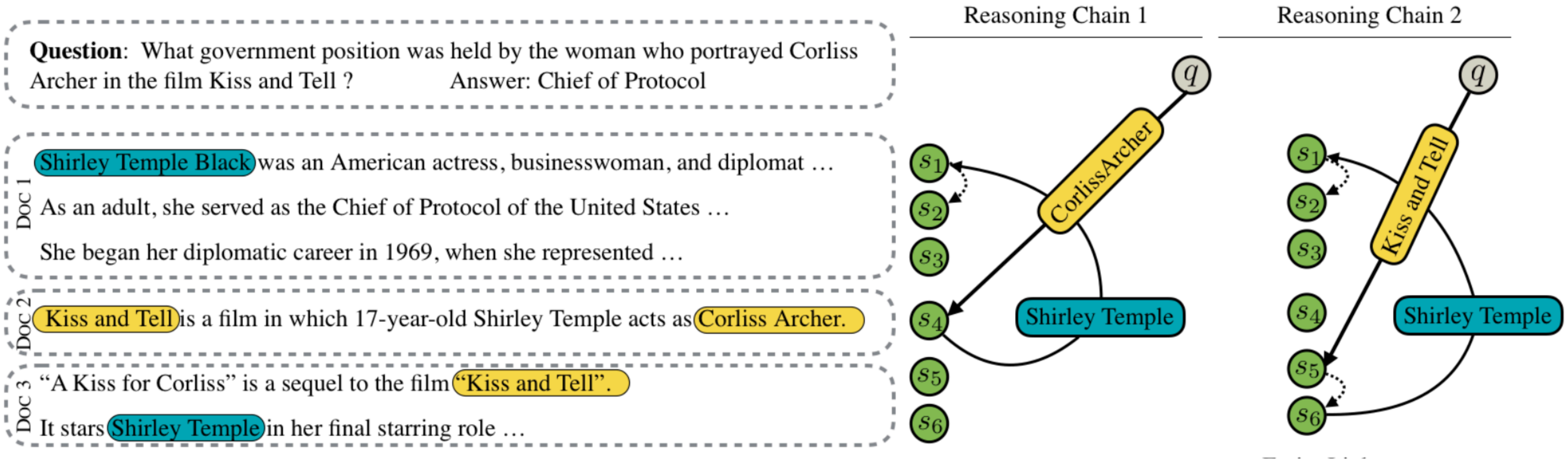
Cons / Future work

(1) Not scale pretty well when we have a large number of operators and constants.

(2) Similar to other models, we still assume the quantity positions are known in advance.

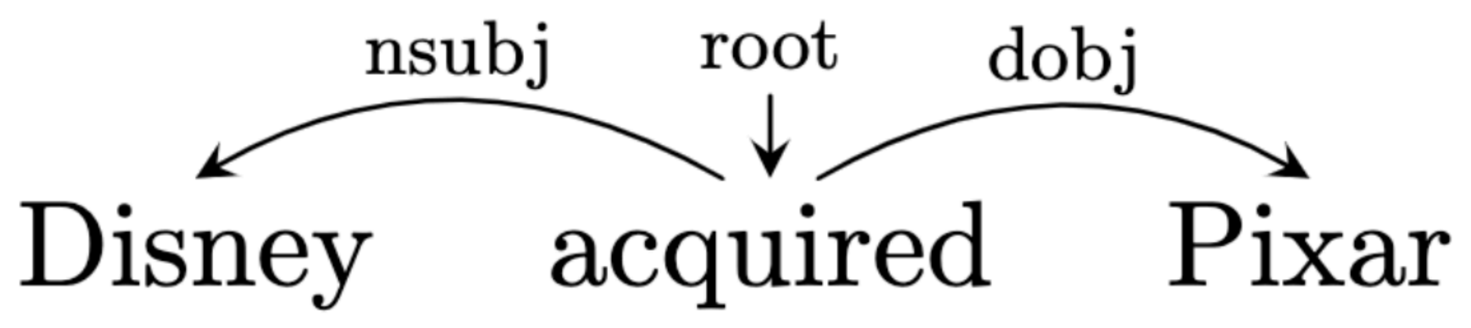
(3) Challenging to apply beam search strategy

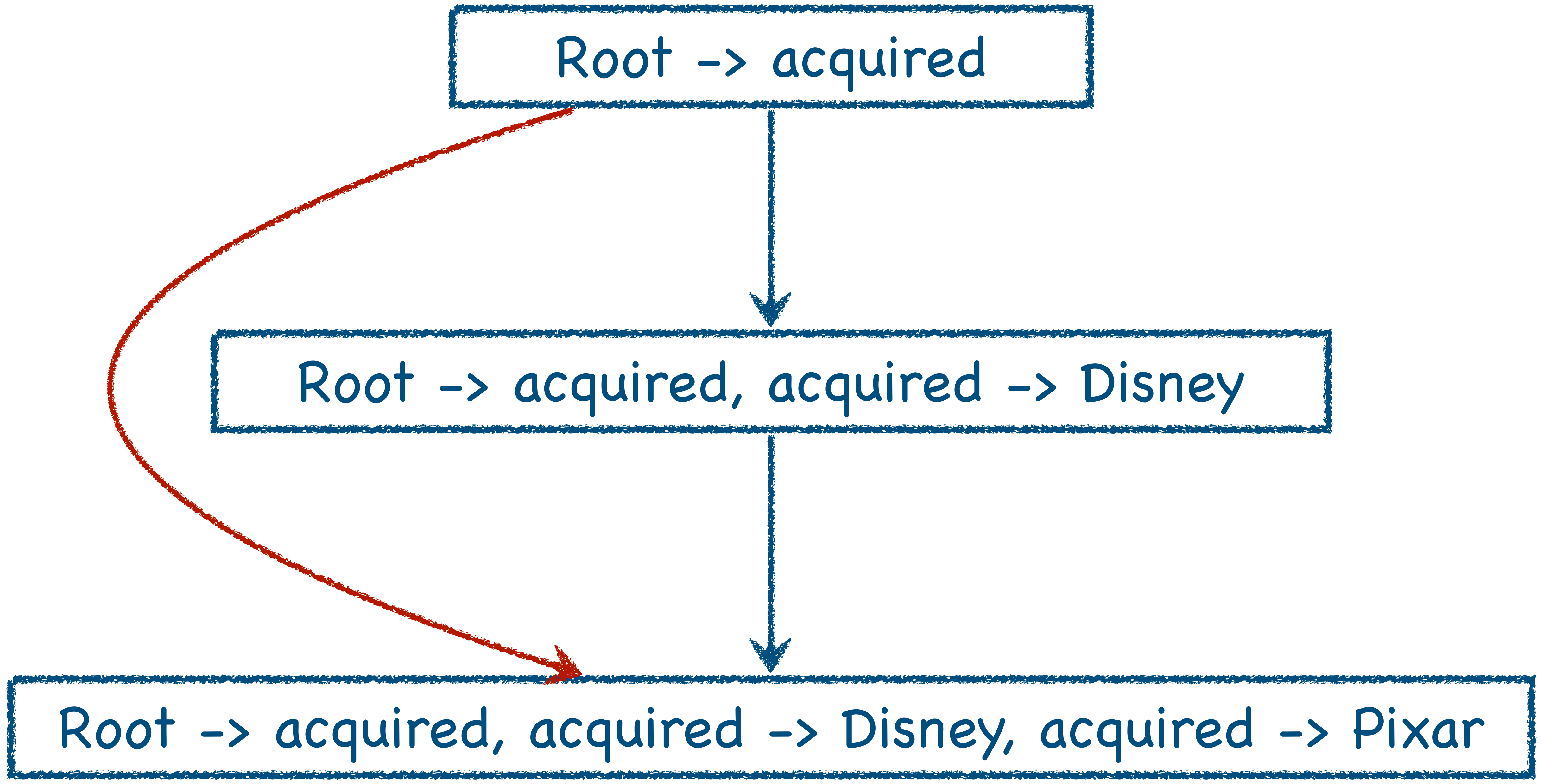
Multi-hop Question Answering via Reasoning Chains (Chen et al., 2021)



Find the reasoning chain among the relevant sentences

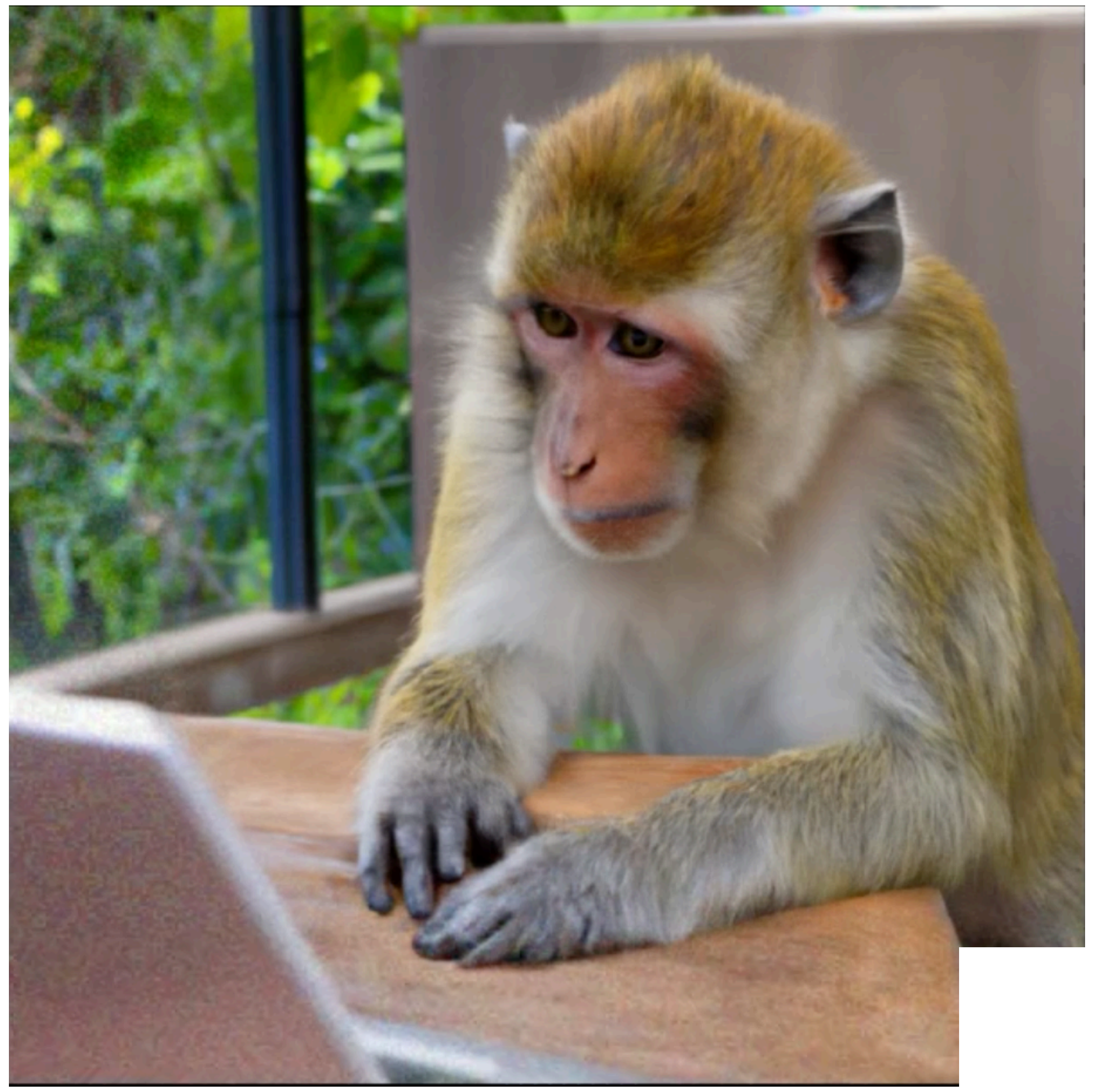
Deductive Systems for Structures

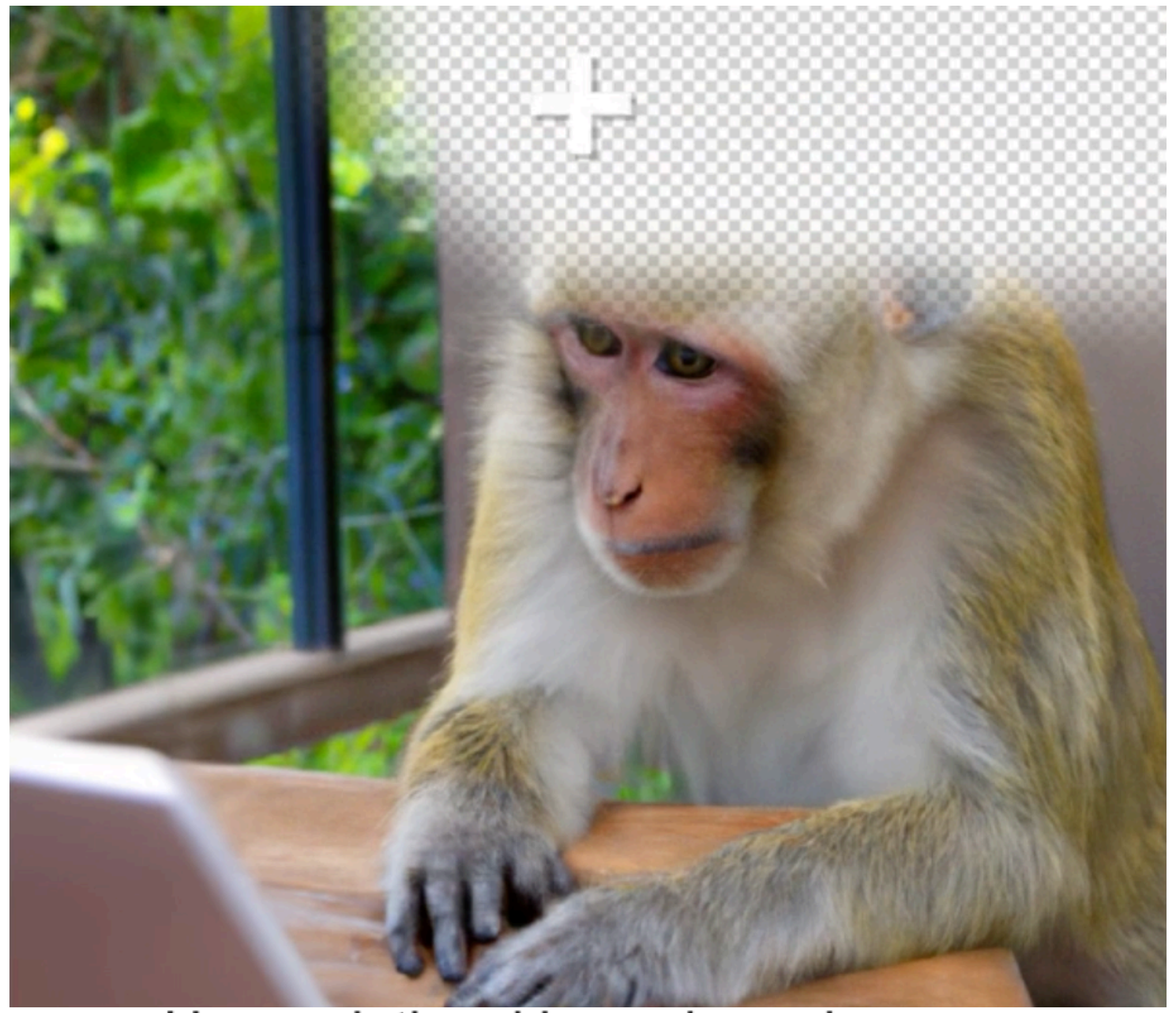




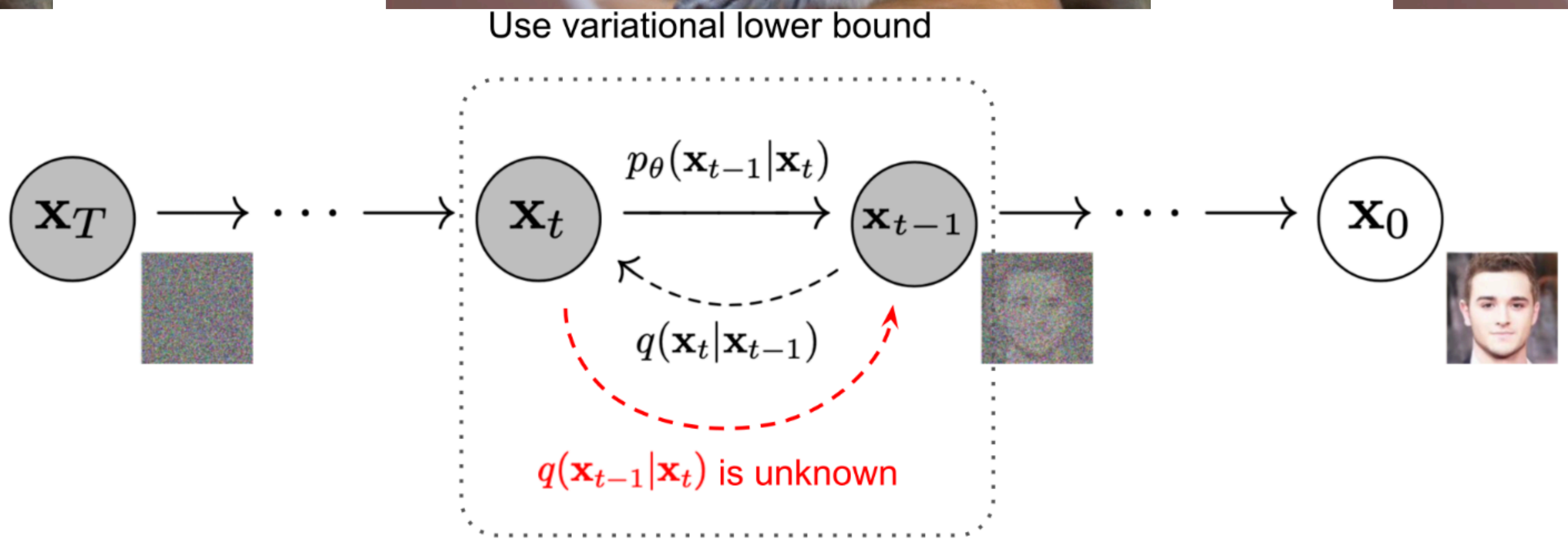
Generating dependency trees in a deductive manner

TLDR: DALL-E-2 by Open AI





https://op



ples

Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. 2020 with a few additional annotations)

TLDR: GPT-3 by OpenAI

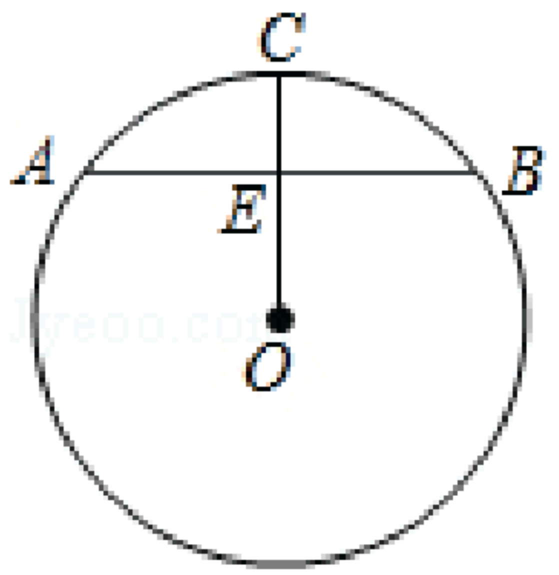
Solving Math Word Problems

WE'VE TRAINED a system that solves grade school math problems with nearly twice the accuracy of a fine-tuned GPT-3 model. It solves about 90% as many problems as real kids: a small sample of 9-12 year olds scored 60% on a test from our dataset, while our system scored 55% on those same problems. This is important because today's AI is still quite weak at commonsense multistep reasoning, which is easy even for grade school kids. We achieved these results by training our model to recognize its mistakes, so that it can try repeatedly until it finds a solution that works.

From: https://openai.com/blog/grade-school-math/

TLDR: Geometric Question

As shown in the figure, in ⊙O, AB is the chord, OC⊥AB, if the radius of ⊙O is 5 (N0) and CE=2 (N1), then the length of AB is ()



A. 2 B. 4 C. 6 D. 8

Answer: D. 8

Problem Type: Length Calculation

Knowledge Points: Vertical Diameter, Pythagorean Theorem

Problem Solving Explanations:

OE=OC-CE=5-2=3. According to the Pythagorean Theorem,

AE = $\sqrt{OA^2 - OE^2}$= $\sqrt{5^2 - 3^2}$= 4. Thus, AB=2AE=8.

Annotated Programs:

Minus | N0 | N1 | PythagoreanMinus | N0 | V0 | Double | V1

Step1: Minus(N0, N1) = 5 – 2 = 3 (V0)

Step2: PythagoreanMinus(N0, V0) = $\sqrt{5^2 - 3^2}$ = 4 (V1)

Step3: Double(V1) = 2×4 = 8 (V2)

From: "GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning"

Solving Differential Equations (Lample and Charton, 2020)

$$162x\log(x)y' + 2y^3\log(x)^2 - 81y\log(x) + 81y = 0 \qquad y = \frac{9\sqrt{x}\sqrt{\frac{1}{\log(x)}}}{\sqrt{c+2x}}$$

	Integration (BWD)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

From: "Deep Learning For Symbolic Mathematics" in ICLR 2020