Guiding Mathematical Reasoning via Mastering Commonsense Formula Knowledge

Jiayu Liu, Zhenya Huang*, Zhiyuan Ma, Qi Liu, Enhong Chen*, Tianhuang Su, Haifeng Liu

University of Science and Technology of China (USTC);
OPPO Mobile Telecommunications;

Outline

Background

Knowledge lays the foundation for human cognition

- Humans naturally acquire knowledge from experience and manipulate it in cognitive behaviors
  - How to gain knowledge from data and apply it in complex reasoning tasks?
- We focus on **Mathematical reasoning** task

algebra

geometry

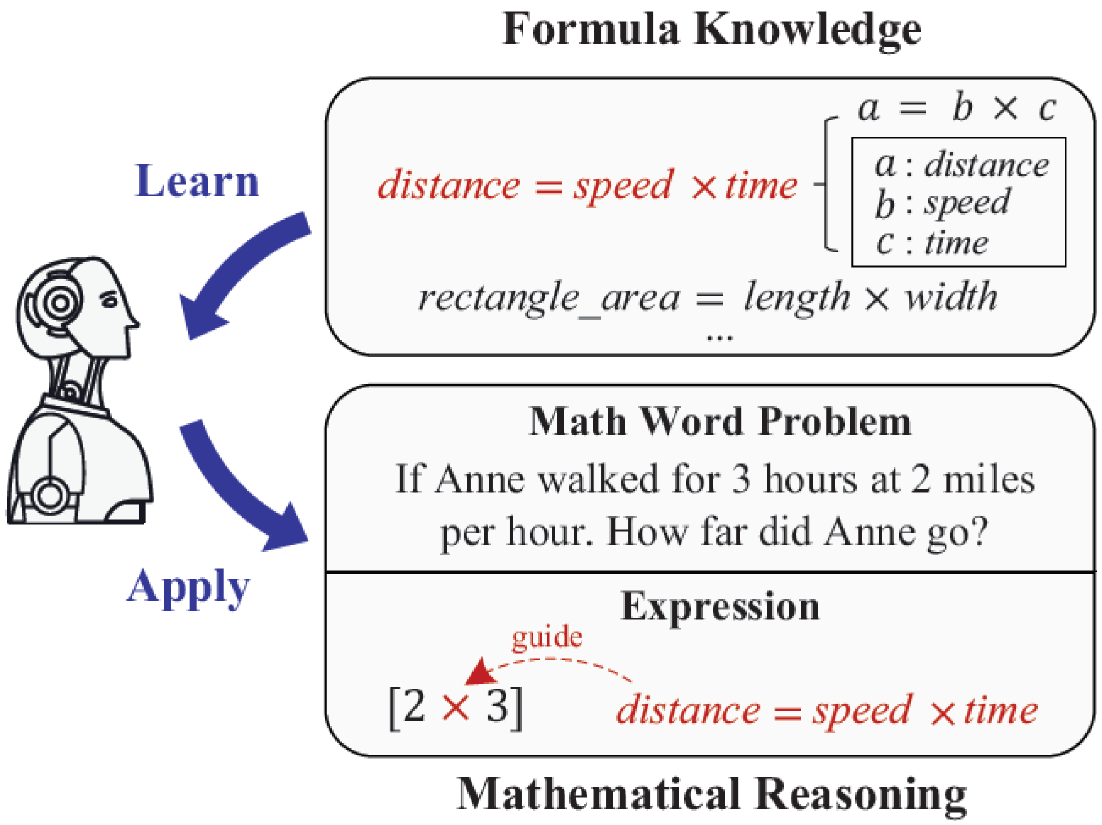
probability

…

Background

Math formulas are essential and commonsense knowledge

- We are constantly **learning** and **applying** various **math formulas**



Math Word Problem (MWP)

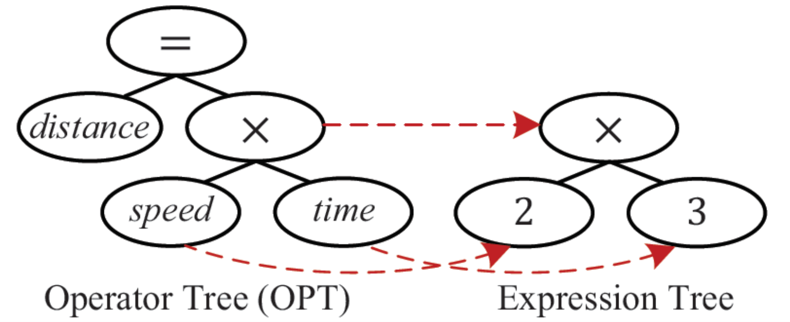
Background

Formula knowledge has not attracted enough attention !

- Existing benchmark datasets do not provide a label for formula usage

  —— E.g, Math23K, MAWPS, GSM8K, MATH

- We **contribute two datasets (Math23K-F and MAWPS-F)**

Problem	If Anne walked for 3 hours at 2 miles per hour. How far did Anne go?
Answer	6
Expression	$[2 \times 3]$
Formula	$[\emptyset, \textit{distance} = \textit{speed} \times \textit{time}, \emptyset]$
Explanation	

Existing data (Problem, Answer, Expression)

Our annotation (Formula, Explanation)

Figure 2: An example of our annotated data.

Background

Formula knowledge has not attracted enough attention !

- We **contribute two datasets (Math23K-F and MAWPS-F)**

More than 25% of SOTA methods' errors are due to the inability to use formulas.

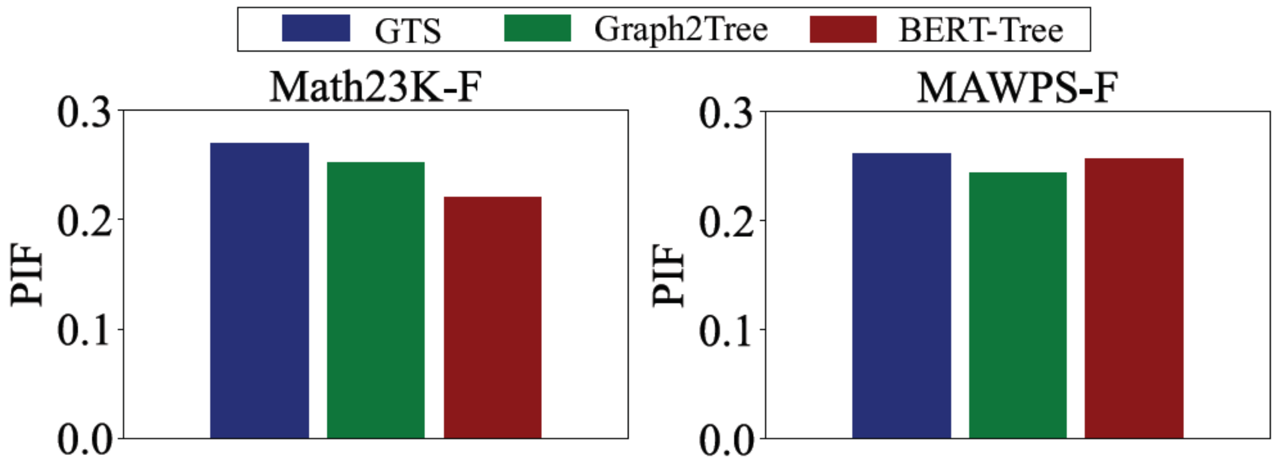


Figure 3: *PIF* of GTS, Graph2Tree, and BERT-Tree.

Other analyses please refer to our paper

Background

Integrating formula knowledge remains many challenges

- Highly symbolic that involve **abstract structure** and **concrete concepts**
  - structure "$a = b \times c$"
  - concepts "$distance, speed, time$"

- Contain rich **mathematical logic**
  - logical transformations such as changing "$distance = speed \times time$" to "$speed = distance \div time$"

- The process of human **application of formula** knowledge is sophisticated

Background

Contributions

- **Dataset:** We **construct two benchmark MWP datasets** (Math23K-F, MAWPS-F) with annotations of the required formula at each reasoning step to support the exploration of formula knowledge in the domain

- **Model:** We **propose a Formula-mastered Solver (FOMAS)** that learns and applies formula knowledge to conduct mathematical reasoning

- **Mechanism:** We design a novel **pretraining manner** to learn the knowledge behind math formulas and develop elaborate **formula-guided mechanisms**. Extensive experiments clearly demonstrate their effectiveness.

Outline

Dataset Construction

- **A valuable dataset should have two characteristics**
  - Preciseness
    - ✓ annotate the formula applied at each reasoning step
  - Generality
    - ✓ most previous models can be easily and fairly compared on them

- **We contribute two datasets: Math23K-F and MAWPS-F:**
  - Annotate the **two most widely studied** MWP datasets: Math23K and MAWPS

Dataset Construction

- **Annotation Process**
  1. Collect essential math formulas from textbooks and summarized 51 and 18 representative formulas on Math23K and MAWPS respectively
  2. Select the most suitable formula for each reasoning step
  3. Evaluate the annotations and repeat evaluation-modification processes

1. Collect formula sets

2. Annotate the formula usage at each step

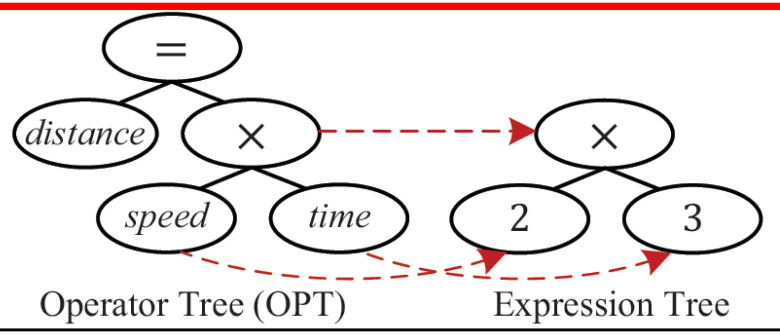
Problem	If Anne walked for 3 hours at 2 miles per hour. How far did Anne go?
Answer	6
Expression	$[2 \times 3]$
Formula	$[\emptyset, distance = speed \times time, \emptyset]$
Explanation	

Figure 2: An example of our annotated data.

3. Evaluation:
Final pass rate is 97.7%

Dataset Analysis

- **Statistics**

Table 1: The 5 most frequently used math formulas.

Math23K-F	1. $distance = speed \times time$
	2. $work = rate \times time$
	3. $total\_cost = unit\_cost \times total\_number$
	4. $total\_amount = unit\_amount \times total\_number$
	5. $total\_weight = unit\_weight \times total\_number$
MAWPS-F	1. $total\_amount = unit\_amount \times total\_number$
	2. $total\_cost = unit\_cost \times total\_number$
	3. $total\_income = unit\_income \times total\_number$
	4. $distance = speed \times time$
	5. $work = rate \times time$

Table 2: Statistics of our benchmark datasets.

Dataset	Math23K-F	MAWPS-F
Num. problems	23,162	2,373
Num. formulas (and variants)	51 (131)	18 (46)
Num. problems requiring formula	7,750	911
Avg. problem length	28.02	30.08
Avg. expr. length	5.55	4.20

Table 8: Distributions of the number of used formulas.

# Used Formulas	# Problems (Math23K-F)	# Problems (MAWPS-F)
0	14,412	1,462
1	4,520	860
2	3,005	33
More than 2	225	18

- **33.5% and 38.4%** of problems require the use of formulas on Math23K-F and MAWPS-F, respectively
- Dataset available at: ***https://github.com/Ljyustc/FOMAS***

Dataset Analysis

- **Whether the study of formula knowledge is necessary?**
  - SOTA MWP models: GTS, Graph2Tree, BERT-Tree
  - PIF metric: $\frac{\text{problems that a model answers incorrectly at steps requiring a formula}}{\text{all problems that it answers incorrectly}}$

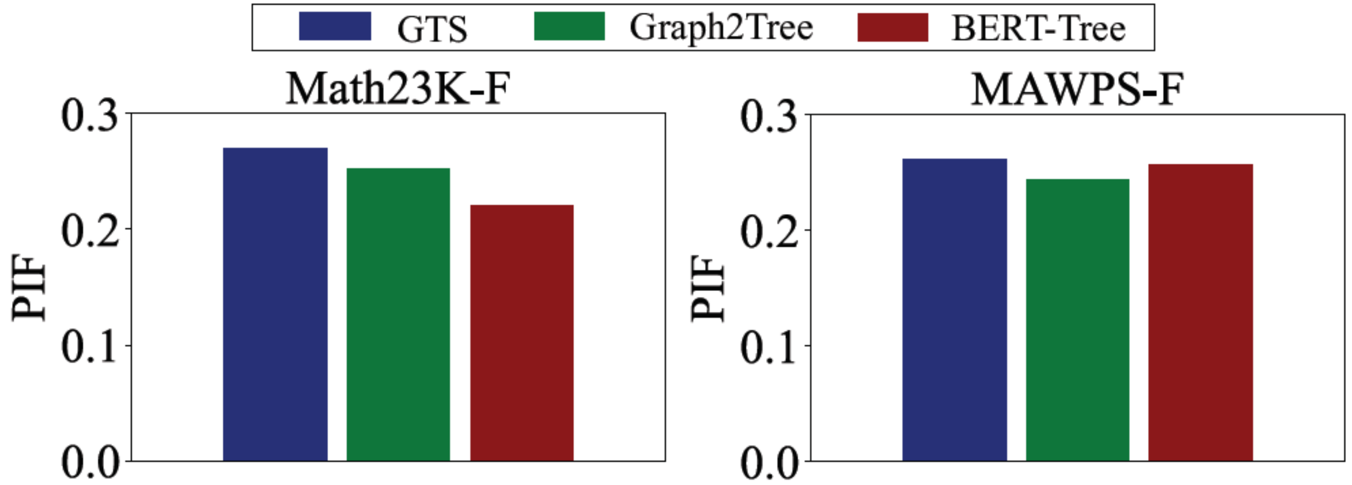


Figure 3: PIF **of GTS, Graph2Tree, and BERT-Tree.**

$\boldsymbol{PIF > 25\%}$

More than 25% of SOTA methods' errors are due to the inability to use formulas.

Outline

Our Method

Problem definition

- Input:
  1. A sequence of n words and numeric values $X_P = [x_1, x_2, \dots, x_n]$
     - ✓ E.g., "If Anne walked for 3 hours … "
  2. **A formula set $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_K\}$**

- Output: mathematical expression Y_P, answer S_P
  - $Y_P = \{y_1, y_2, \dots, y_m\}$, where y_i comes from $V_P = V_O \cup V_C \cup N_P$
    - ✓ V_O: operators, e.g. $\{+, \times, -, \div\}$
    - ✓ V_C: numeric constants, e.g. $\{1, \pi\}$
    - ✓ N_P: numeric variables from P, e.g. $\{3,2\}$
    - ✓ E.g., "3×2"
  - S_P: real value
    - ✓ E.g., 6

Problem	If Anne walked for 3 hours at 2 miles per hour. How far did Anne go?
Answer	6
Expression	$[2 \times 3]$
Formula	$[\emptyset, \textit{distance} = \textit{speed} \times \textit{time}, \emptyset]$

Our Method

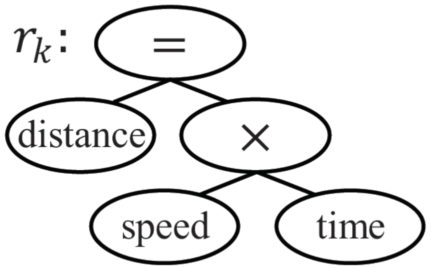
Problem

Table 2: Statistics of our benchmark datasets.

Dataset	Math23K-F	MAWPS-F
Num. problems	23,162	2,373
Num. formulas (and variants)	51 (131)	18 (46)
Num. problems requiring formula	7,750	911
Avg. problem length	28.02	30.08
Avg. expr. length	5.55	4.20

- Formula Knowledge
  - Set of math formulas $R = \{r_1, r_2, \dots, r_K\}$
  - Formula $r_k = [z_1, z_2, \dots, z_l]$ to store the **prefix expression of its Operator Tree (OPT)**
    - ✓ Note: "=" must be the root of OPT
    - ✓ E.g., $[=, \boldsymbol{distance}, \times, \boldsymbol{speed}, \boldsymbol{time}]$) ← **Distance = speed ×time**
  - Each z_i:
    - ✓ Abstract concept (e.g., "distance")
    - ✓ Operator (e.g., "=","×")
  - For each formula $r_k \in R$, define a set of variants $A(r_k) \nsubseteq R$
    - ✓ Imply Mathematical transformations
    - ✓ E.g., "$speed = distance \div time$" and "$time = distance \div speed$"

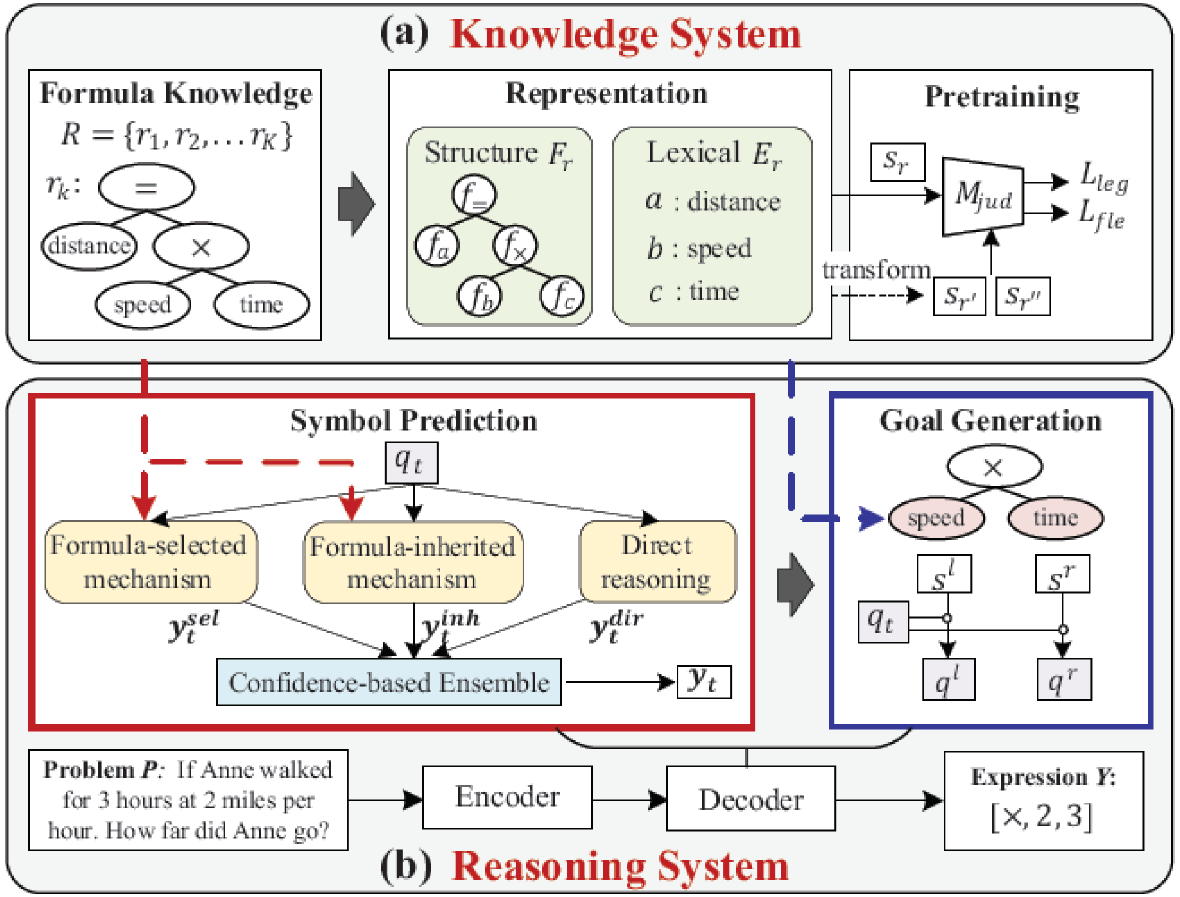
Operator Tree (OPT)



Our Method

FOMAS: Formula-mastered Solver

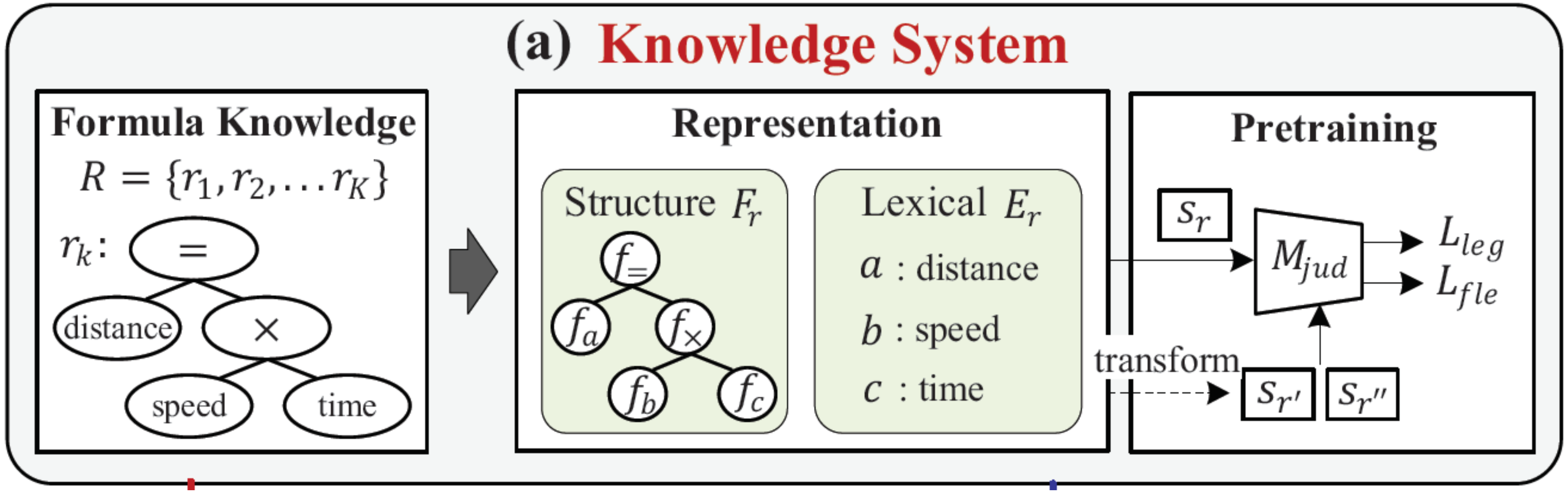
- Idea: inspired by dual process theory, we construct two systems: **Knowledge System** and **Reasoning System**
  - **Knowledge System**: store formula knowledge and mimic how humans represent and learn it
  - **Reasoning System**: conduct mathematical reasoning by applying the formula knowledge in Knowledge System



Our Method

● Knowledge System: Learn formula knowledge

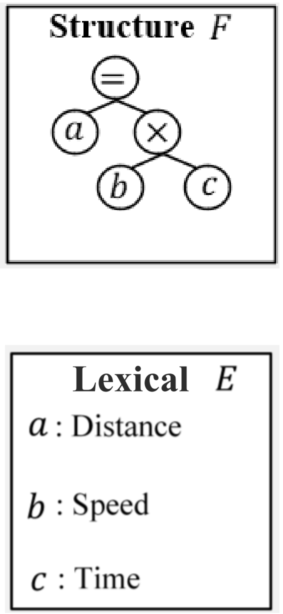
- **Formula Representation**: mine the features of formulas
- **Formula Pretraining**: learn the features of formulas by resembles humans' mastery of the mathematical logic behind them

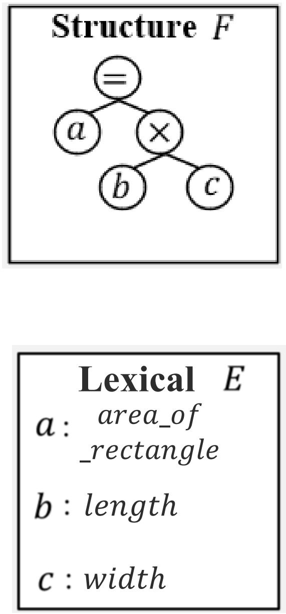


Knowledge System

- Formula Representation
  - Each formula r is compositionally built up from the combination of two types of information：
    - ✓ **Structural Feature F_r**： defines the order and mode of formula calculation
    - ✓ **Lexical Feature E_r**： refers to the concrete meanings

$distance = speed \times time$

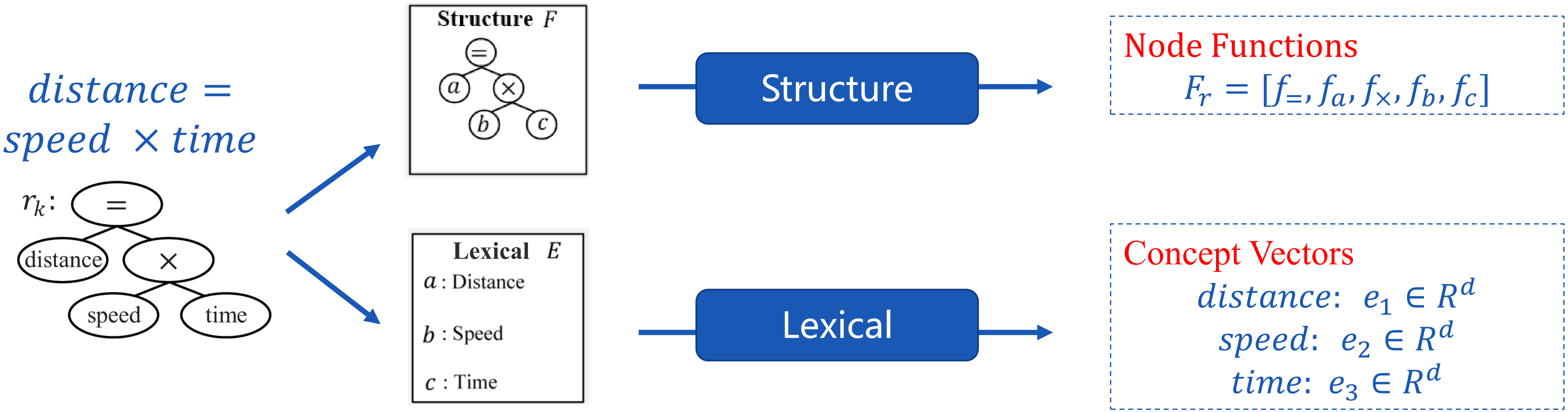




$area\_of\_rectangle = length \times width$

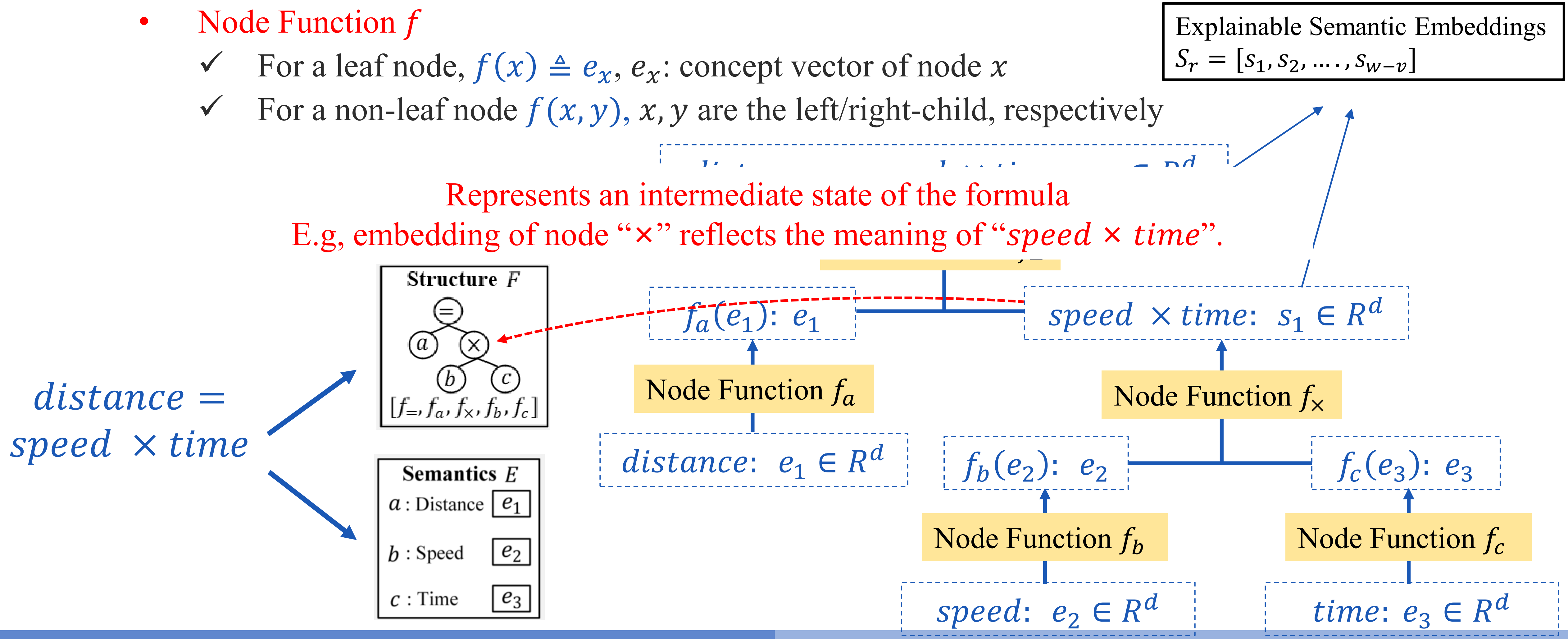
Knowledge System

- Formula Representation
  - Each formula r is compositionally built up from the combination of two types of information：
    - ✓ **Structural Feature F_r**： defines the order and mode of formula calculation
    - ✓ **Lexical Feature E_r**： refers to the concrete meanings



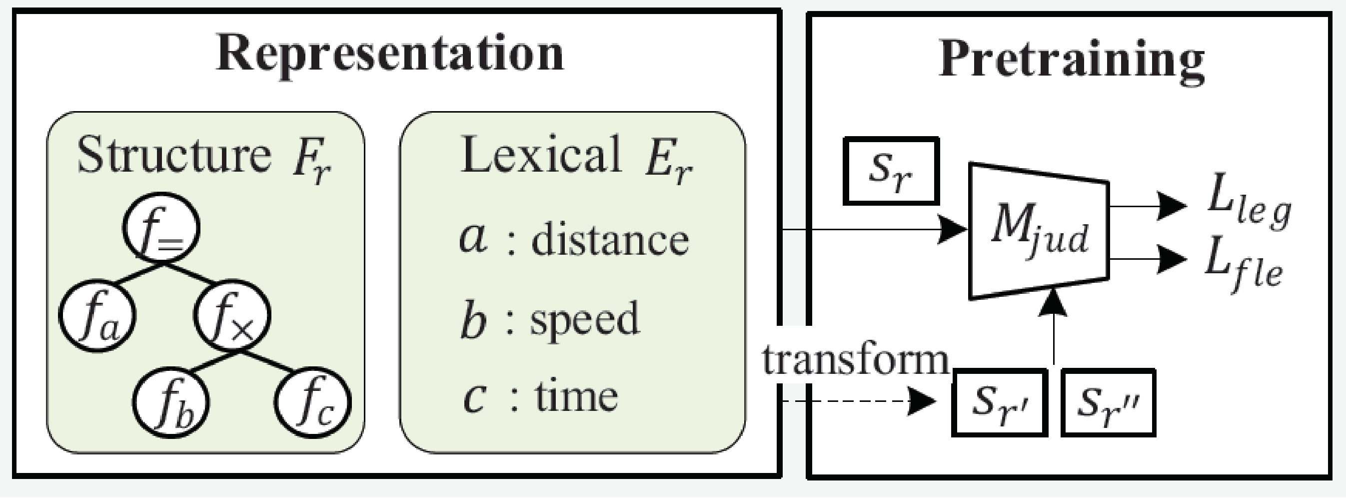
Knowledge System

- Formula Representation
  - Node Function f
    - ✓ For a leaf node, $f(x) \triangleq e_x$, e_x: concept vector of node x
    - ✓ For a non-leaf node $f(x, y)$, x, y are the left/right-child, respectively



Knowledge System

- Formula Pretraining
  - Mimicking how humans process and learn the formulas autonomously
  - Pretrain: node functions $f_i \in F_r$, concept vectors in E_r
  - Two objective: **Legality** L_{leg}, **Flexibility** L_{fle}

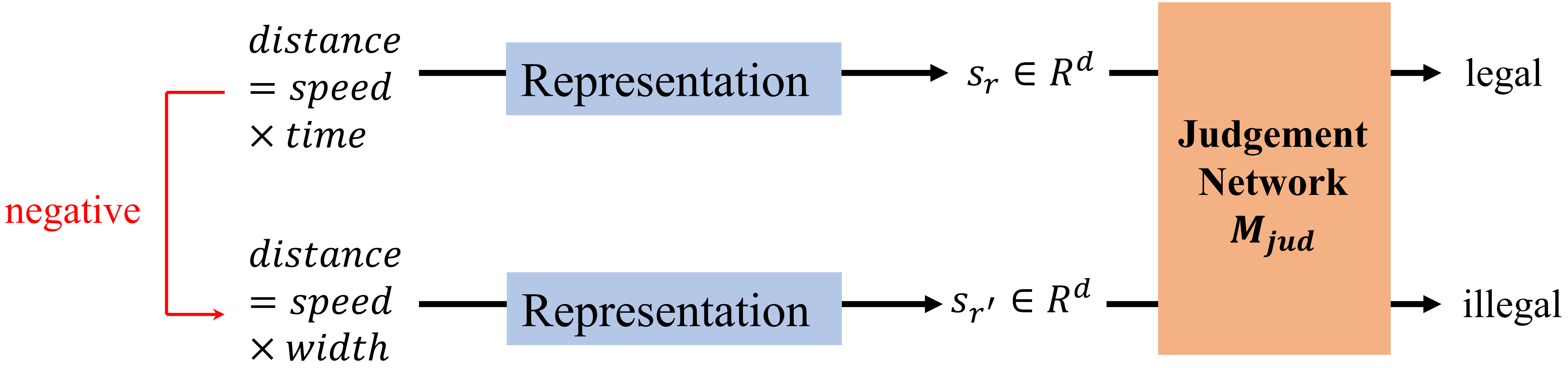


Legality L_{leg}

Flexibility L_{fle}

Knowledge System

- Formula Pretraining
  - **Legality**: understanding whether a formula is legal or not
    - ✓ E.g., $distance = speed \times time$ ✓ $distance = speed \times width$ ✗

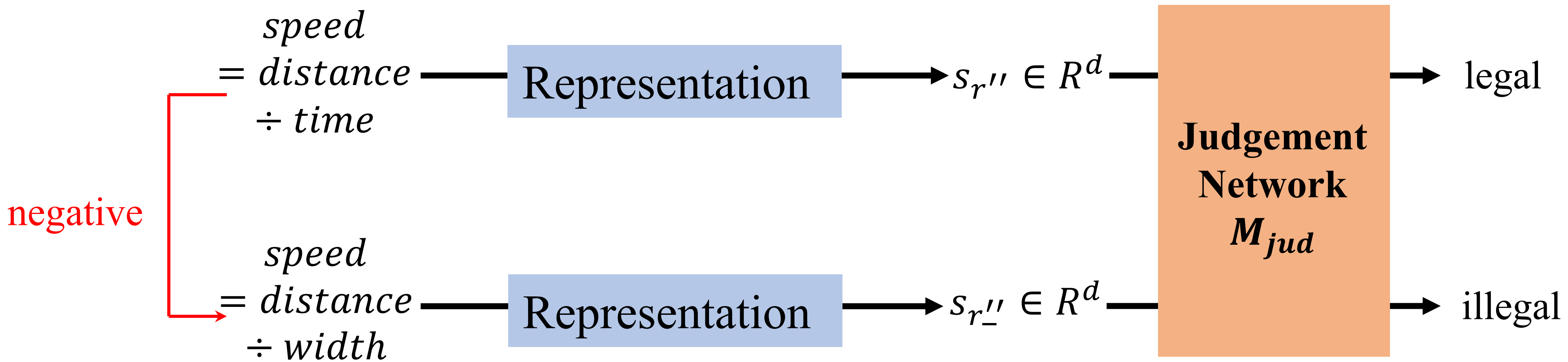


$$L_{leg} = \sum_{r \in R} l_{BCE}(M_{jud}(\mathbf{s}_r), 1) + l_{BCE}(M_{jud}(\mathbf{s}_{r'}), 0),$$

$$M_{jud}(\mathbf{s}) = \sigma(\mathbf{W}_{u1} \cdot ReLu(\mathbf{W}_{u2} \cdot tanh(\mathbf{W}_{u3} \cdot \mathbf{s}))),$$

Knowledge System

- Formula Pretraining
  - **Flexibility**: understanding the complex transformation of formulas
    - ✓ E.g., $distance = speed \times time$ ✓ $speed = distance \div time$ ✓



$$L_{fle} = \sum_{r \in R} \sum_{r'' \in A(r)} l_{BCE}(M_{jud}(s_{r''}), 1) + l_{BCE}(M_{jud}(s_{r''_-}), 0).$$

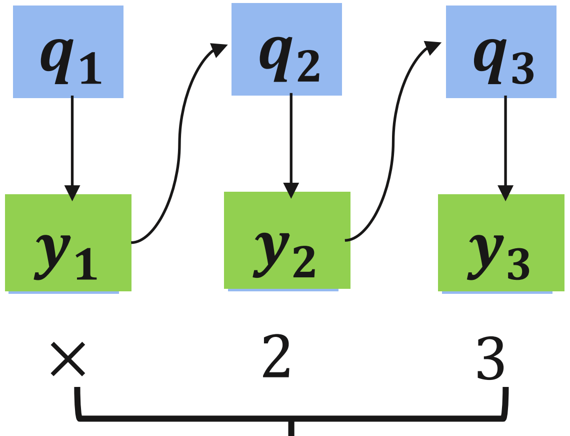
$$L_{pretrain} = L_{leg} + L_{fle}.$$

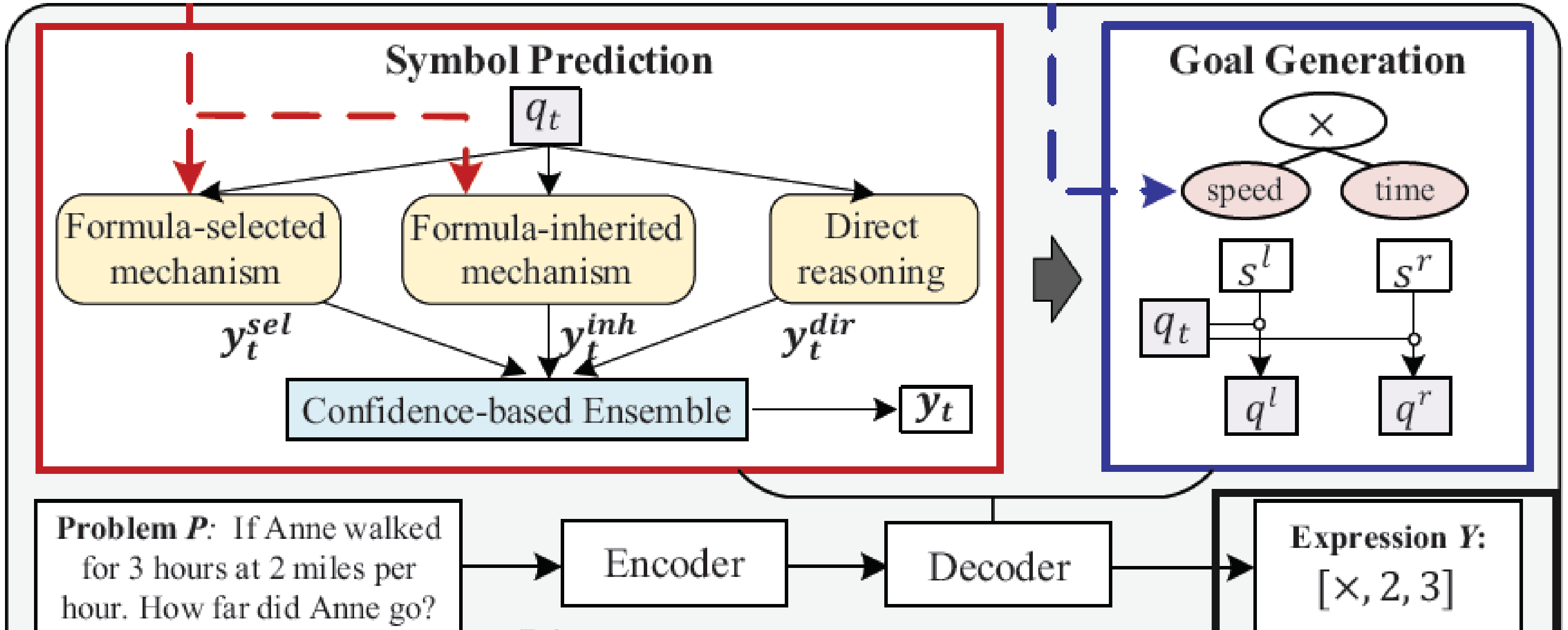
Reasoning System

- Encoder-Decoder based
  1. **predicts the symbol y_t** given the reasoning goal q_t — $q_t \rightarrow y_t$
  2. **generates the next reasoning goal** q_{t+1} — $q_t, y_t \rightarrow q_{t+1}$
- Formula knowledge guides both **symbol prediction** and **goal generation** in decoder

Training Loss:

$$L = \sum_{P} \sum_{t} -log\ P(y_t | y_1, \ldots, y_{t-1}, \boldsymbol{q}_t, \boldsymbol{H})$$





(b) Reasoning System

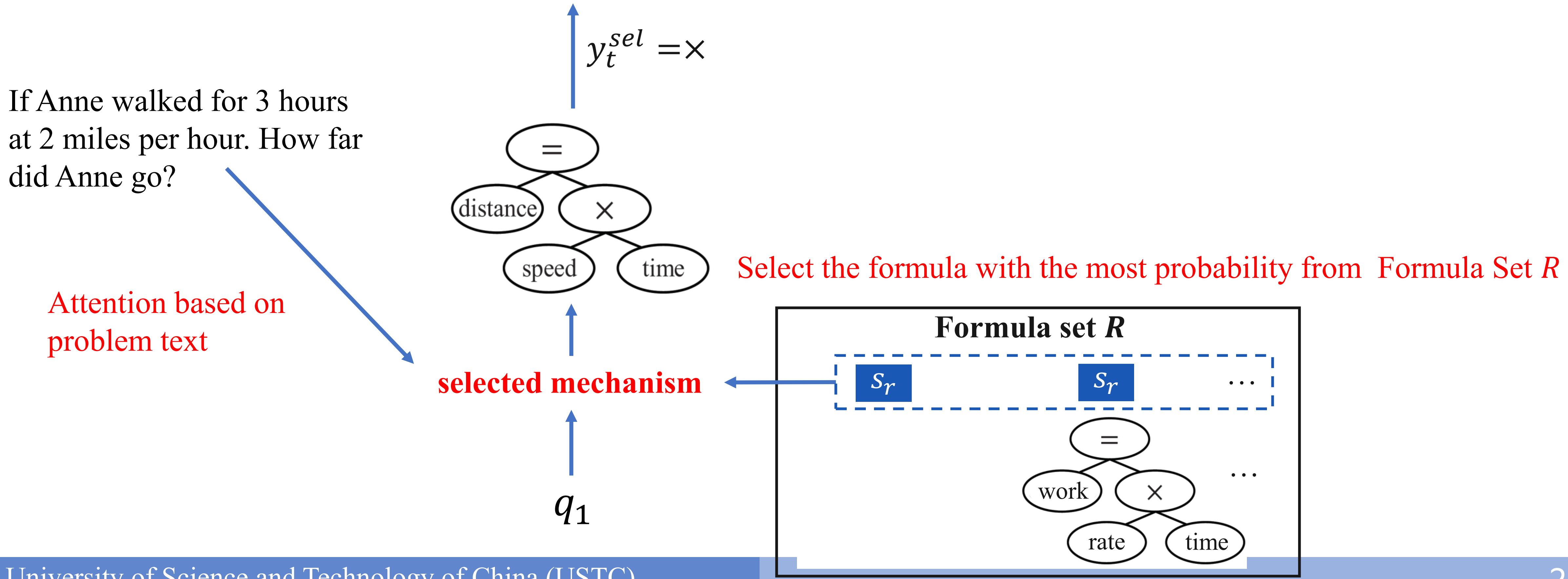
Reasoning System

- **Formula-guided symbol prediction:** $q_t \rightarrow y_t$
  - Propose **three** types of symbolic reasoning mechanisms summarized from sophisticated human thought process
    - ✓ 1. Formula-selected mechanism
    - ✓ 2. Formula-inherited mechanism
    - ✓ 3. Direct reasoning
  - Confidence-based Ensemble

Reasoning System

1. Formula-selected mechanism
   - **Idea**: Retrieve a formula from **Knowledge System** and extract a symbol from it

$$E = [\times, \quad 2, \quad 3]$$

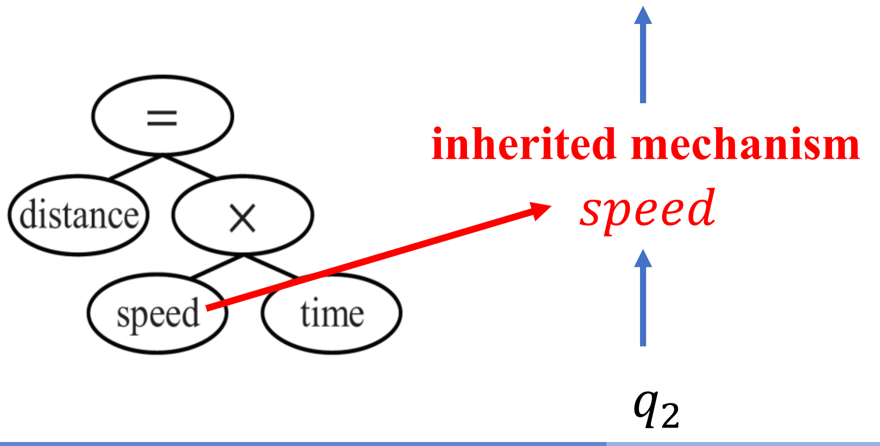


Reasoning System

2. Formula-inherited mechanism
   - **Idea**: The selected formula implies a kind of **thinking pattern** that navigates **multiple** reasoning steps
   - **Implementation**:
     - ✓ if the descendant is an operator c, $y_t^{inh} = one - hot(o_c)$
     - ✓ if the descendant is leaf node (e.g.,"*speed*") $y_t^{inh} = softmax(\mathbf{W}_{h1} \cdot tanh(\mathbf{W}_{h2} \cdot [\mathbf{e}_{inh}, \mathbf{q}_t]))$.

$E = [\times, \quad 2, \quad 3]$

If Anne walked for 3 hours at 2 miles per hour. How far did Anne go?



Reasoning System

3. Direct reasoning mechanism
   - Directly reason a symbol y_t^{dir} without any formula, implemented as GTS

- Confidence-based Ensemble
  - **Idea**: Inspired by Mixture of Experts (MoE)
    - ✓ $\boldsymbol{P_{sel}}, \boldsymbol{P_{inh}}$ reflects the confidence of using the selected/inherited formula

$$y_t = \frac{P_{sel}}{P_{sel} + P_{inh} + 1} \cdot y_t^{sel} + \frac{P_{inh}}{P_{sel} + P_{inh} + 1} \cdot y_t^{inh} + \frac{1}{P_{sel} + P_{inh} + 1} \cdot y_t^{dir}$$

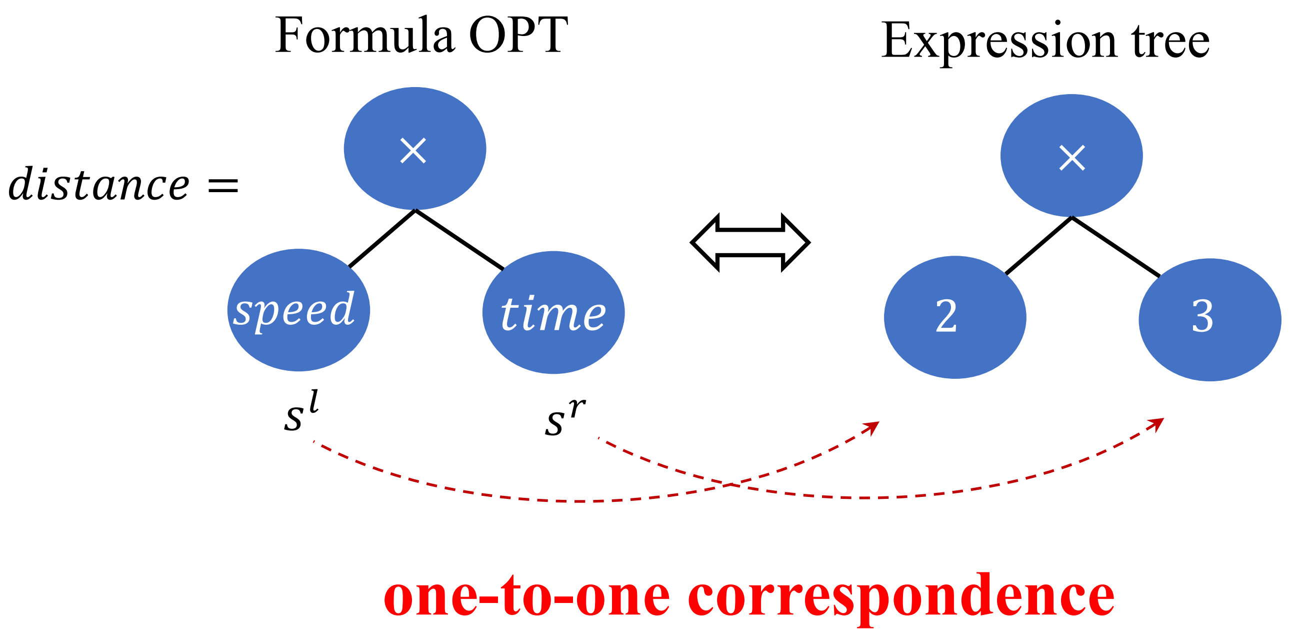
$E = [\times, 2, 3]$

    - ✓ Specially, **at $t = 1$**, no formula-inherited mechanism

$$y_1 = \frac{P_{sel}}{P_{sel} + P_{inh} + 1} \cdot y_t^{sel} + \frac{P_{inh}}{P_{sel} + P_{inh} + 1} \cdot y_t^{inh} + \frac{1}{P_{sel} + P_{inh} + 1} \cdot y_t^{dir}$$

Reasoning System

- **Formula-guided goal generation:** $q_t, y_t \rightarrow q_{t+1}$
  - **Idea**: utilize the concept vector (or semantic embedding) of y_t's **left/right child** in OPT to guide generate the **left/right sub-goal** q^l (i.e., q_{t+1}) / q^r of y_t respectively.



$$q^l = o^l \odot d^l,$$

$$o^l = \sigma(W_o \cdot [s^l, q_t, c]),\ d^l = tanh(W_d \cdot [s^l, q_t, c])$$

Outline

Experiment

Setups

- Dataset
  - Math23K-F
  - MAWPS-F
- Baseline methods
  - Seq2Seq (2015)
  - GTS (2019)
  - Graph2Tree (2020)
  - HMS (2021)
  - NS-Solver (2021)
  - BERT-Tree (2022)
  - SUMC (2022)
  - LogicSolver (2022)
  - ChatGPT (2022)

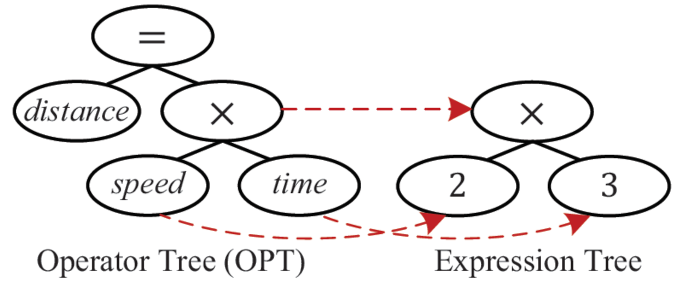
Problem	If Anne walked for 3 hours at 2 miles per hour. How far did Anne go?
Answer	6
Expression	$[2 \times 3]$
Formula	$[\emptyset, \mathit{distance} = \mathit{speed} \times \mathit{time}, \emptyset]$
Explanation	

Figure 2: An example of our annotated data.

Table 2: Statistics of our benchmark datasets.

Dataset	Math23K-F	MAWPS-F
Num. problems	23,162	2,373
Num. formulas (and variants)	51 (131)	18 (46)
Num. problems requiring formula	7,750	911
Avg. problem length	28.02	30.08
Avg. expr. length	5.55	4.20

Experiment

Overall Accuracy

- ✓ **Mastering the formula knowledge** is necessary and valuable to achieve stronger mathematical reasoning ability
- ✓ **Conducting reasoning under the instruction of formula** is necessary and more in line with human cognitive process
- ✓ Capability to acquire and explicitly apply formula knowledge makes our FOMAS **more robust** in figuring out complex problems

Table 3: Answer Accuracy ($* : p < 0.05$ w.r.t. BERT-Tree).

	Math23K-F	MAWPS-F
Seq2Seq	0.640	0.797
GTS	0.756	0.826
Graph2Tree	0.774	0.837
HMS	0.761	0.803
NS-Solver	0.757	/
BERT-Tree	0.833	0.872
SUMC	0.825	0.820
LogicSolver	0.834	/
ChatGPT	0.649	0.883
FOMAS	**0.848****	**0.886***

Experiment

Ablation Study

- ✓ **<u>All components</u>** of FOMAS contribute
- ✓ The removal of **<u>legality or flexibility</u>** has the **<u>greatest impact</u>** on the effect
- ✓ Formula-guided **<u>symbol prediction contributes more</u>** to FOMAS than goal generation

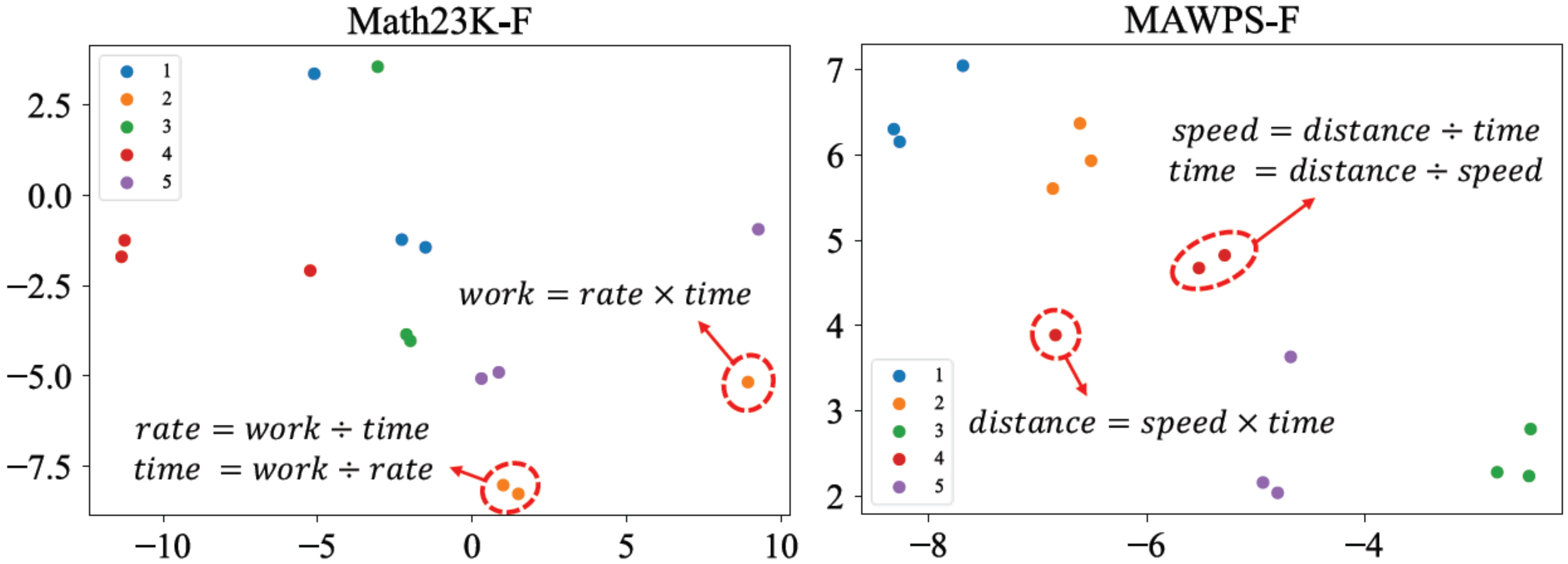
Table 4: Results of ablation study.

		Math23K-F	MAWPS-F
FOMAS		**0.848**	**0.886**
Knowledge System	w/o legality	0.829	0.875
	w/o flexibility	0.832	0.875
Reasoning System	w/o select	0.839	0.878
	w/o inherit	0.843	0.880
	w/o formula-goal	0.842	0.882

Experiment

Formula Learning

- ✓ Learning: investigating the distribution of different formulas' representations



- ✓ **Advantage** in formula learning through **encoding the structural and lexical information separately**
- ✓ Our pretraining manner **grasps the knowledge** behind mathematical transformations

Experiment

Formula Applying

- ✓ Applying: the performance of FOMAS' formula-selected mechanism

	Math23K-F		MAWPS-F	
	FOMAS	IP	FOMAS	IP
ACC(↑)	0.954	0.950	0.961	0.957
Precision(↑)	0.749	0.731	0.808	0.776
Recall(↑)	0.687	0.621	0.758	0.742
	FOMAS	BT	FOMAS	BT
PIF(↓)	0.112	0.220	0.134	0.257

Baselines:
IP: Inner Product
BT: BERT-Tree

- ✓ **Effectiveness and robustness** of formula-selected mechanism
- ✓ FOMAS significantly **reduces the proportion of errors** caused by inability use of formulas
- ✓ FOMAS performs better at formulas that occur a lot

Experiment

Interpretability Verification

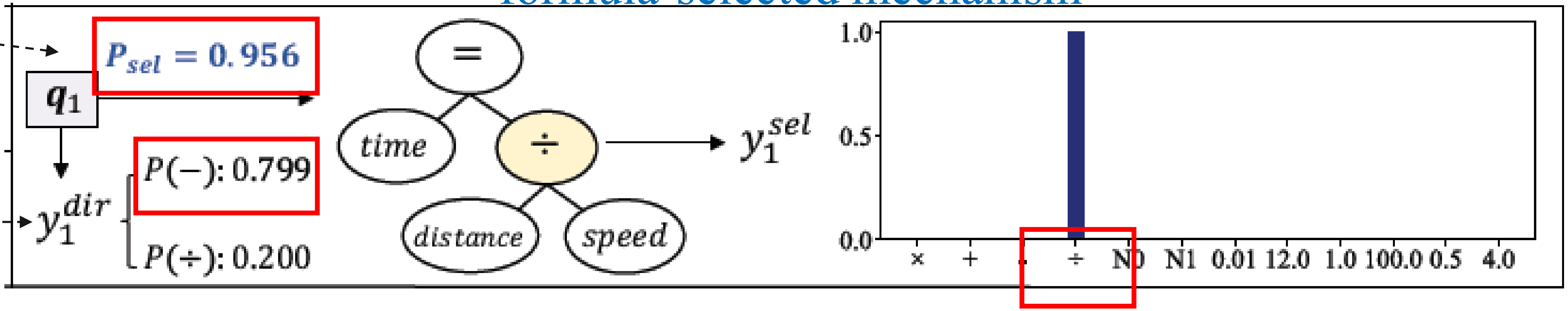
More cases are presented in our paper

Problem 1: If Benjamin skated $N0$ kilometers at $N1$ kilometers per hour , how long was Benjamin skating ?	BERT-Tree $[-,\ N0,\ N1]$ ✗
	FOMAS $[\div,\ N0,\ N1]$ ✓

formula-selected mechanism

Our FOMAS

BERT-Tree



Experiment

Analysis of ChatGPT

Table 3: Answer Accuracy (∗ : $p < 0.05$ w.r.t. BERT-Tree).

	Math23K-F	MAWPS-F
ChatGPT	0.649	0.883
FOMAS	**0.848***	**0.886***

✓ Implementation: on January, 2023; obtain ChatGPT's responses by a crawler; manually extract the numeric answer

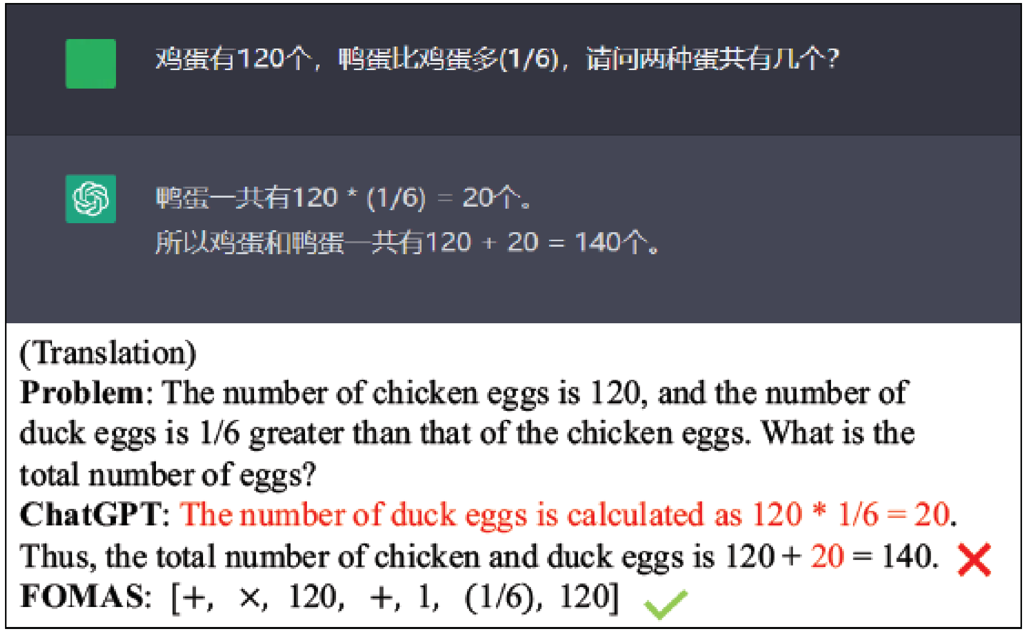


Figure 11: A fraction case between ChatGPT and FOMAS

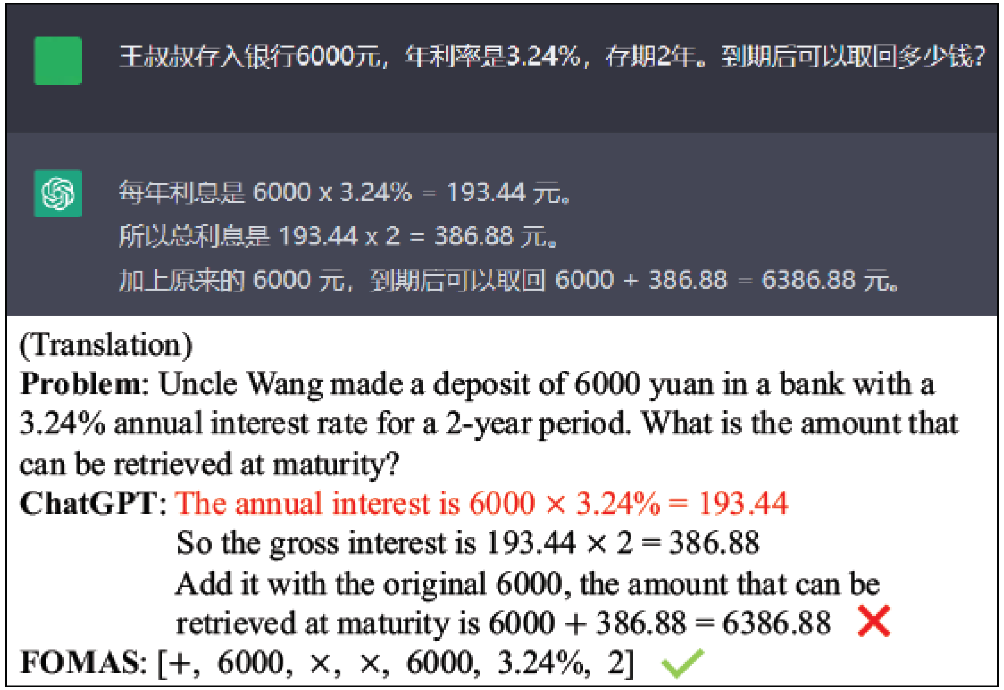
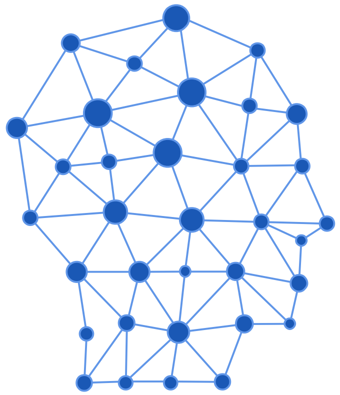


Figure 12: An arithemetic case between ChatGPT and FOMAS

Outline



Conclusion

Summary

- Constructed two benchmark datasets named Math23K-F and MAWPS-F
- Formula-mastered Solver(FOMAS) for math formula learning and applying
    - contained Knowledge-Reasoning Systems inspired by human cognitive structure
    - elaborate formula learning/applying mechanisms
- Experimental results proved the effectiveness and interpretability

Future Work

- Acquire more types of symbolic knowledge from data automatically
- Generalize to more datasets
- …

Thanks for your listening!

For more details, please refer to our paper

Welcome to discuss with us

Jiayu Liu
jy251198@mail.ustc.edu.cn

https://github.com/Ljyustc/FOMAS

中国科学技术大学
University of Science and Technology of China