

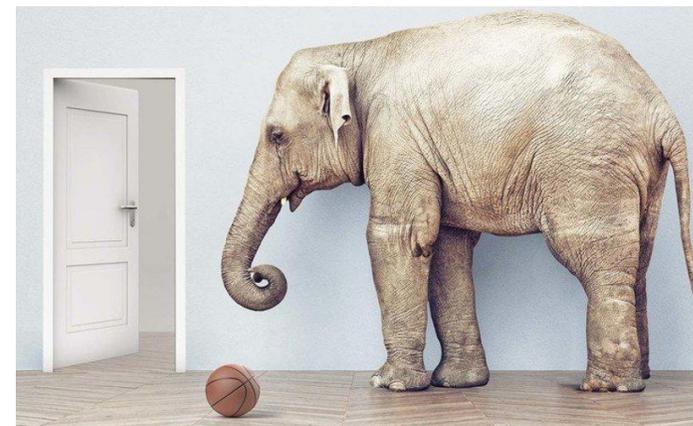
UCLA CS 263: Natural Language Processing

Spring 2023, Instructor: Prof. Kai-Wei Chang

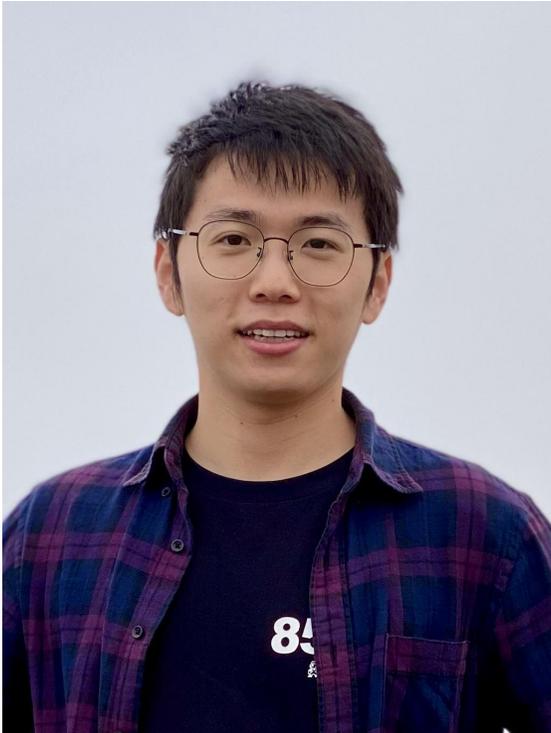
UCLA

Mathematical Reasoning and Commonsense Reasoning

Pan Lu
2023.05.31



About Me



<https://lupantech.github.io/>

- 4th-year Ph.D. Candidate at UCLA
- Research interests
 - ❖ **Large language models** for planning, reasoning, and generation
 - ❖ **Mathematical reasoning** in mathematics and sciences domains
 - ❖ **Trustworthy NLP**: explainable, reliable, and socially responsible
 - ❖ **Conversational agents**: value-aware and socially intelligent
 - ❖ **Multimodal reasoning** for vision-and-language applications
- Workshops and tutorials
 - ❖ Lead organizer for NeurIPS-21 MathAI4ED Workshop
 - ❖ Lead organizer for NeurIPS-22 MathAI Workshop
 - ❖ Keynote presenter at IJCAI-23 Tutorial on math reasoning

Outline

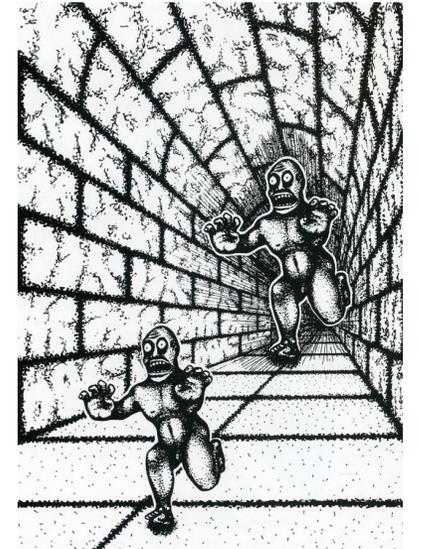
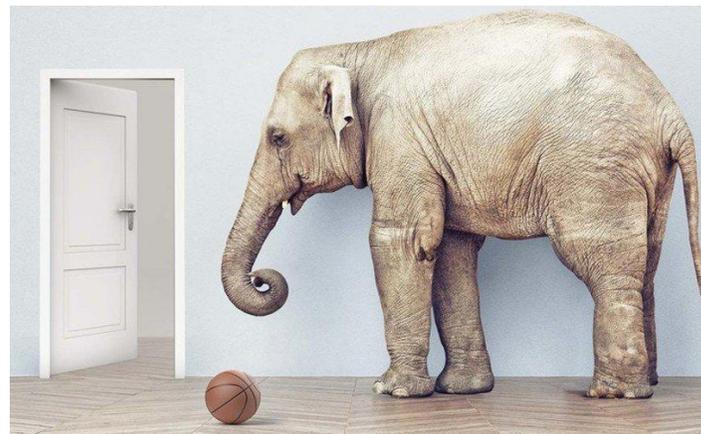
- **Mathematical Reasoning**

- Tasks and benchmarks
- Neural network methods
- Language models and LLMs
- Challenges and opportunities



- **Commonsense Reasoning**

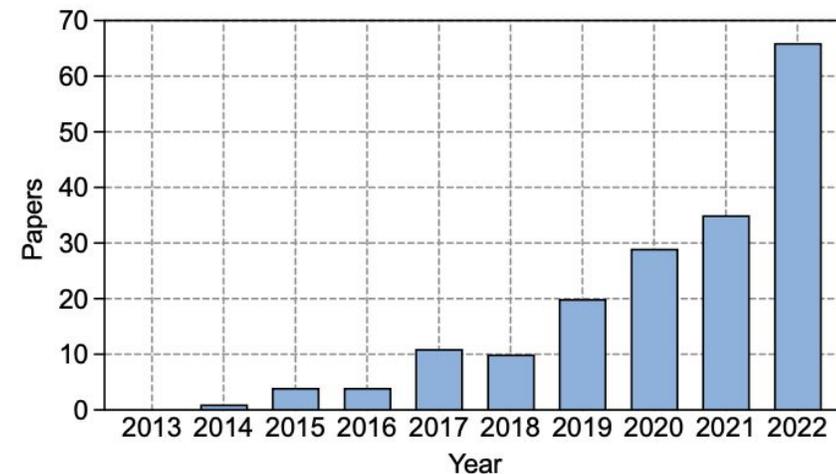
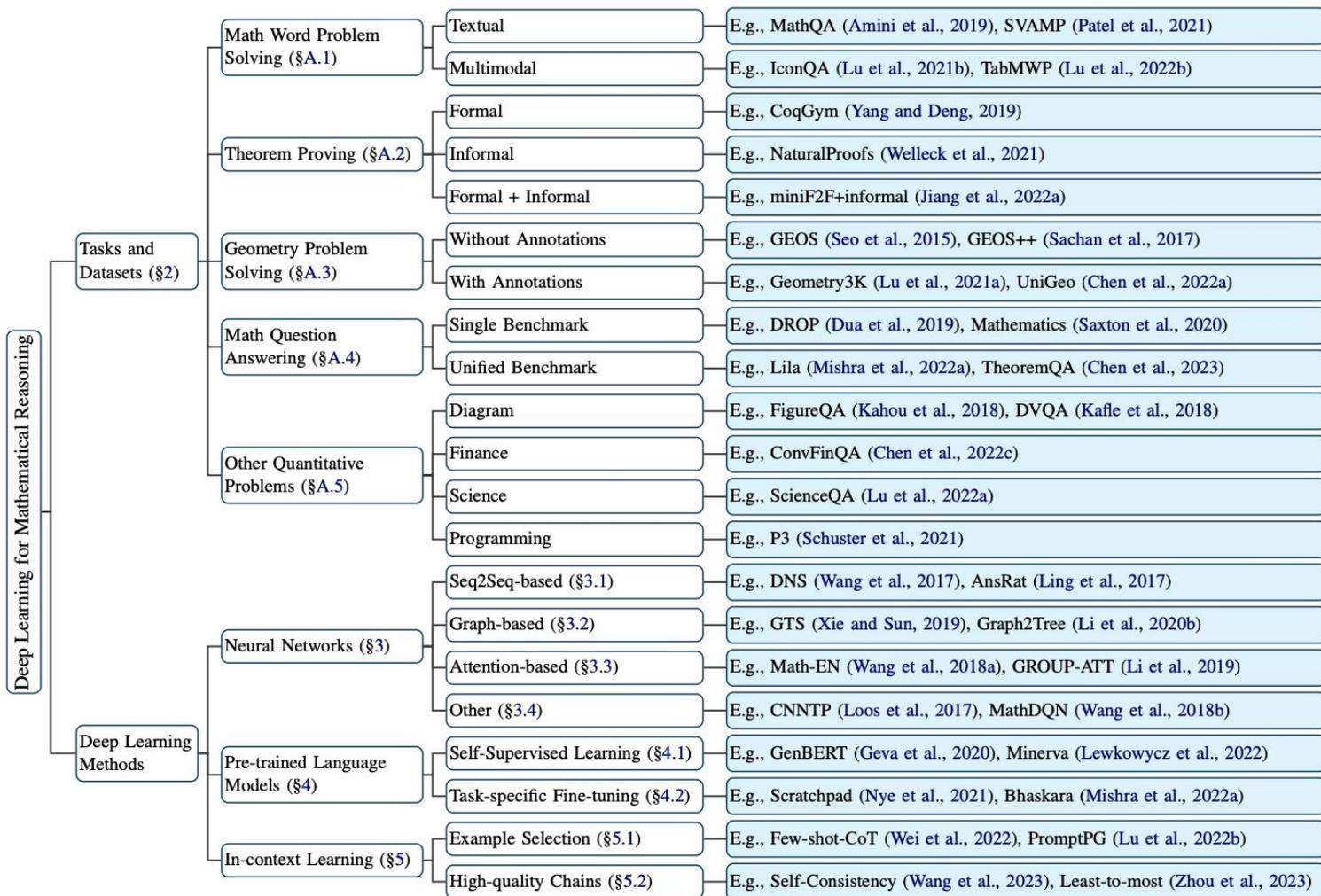
- Definition of commonsense
- Commonsense benchmarks
- Commonsense integration into NNs





*Mathematical
Reasoning*

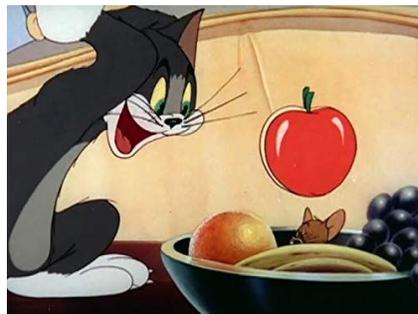
Deep Learning for Mathematical Reasoning



MWP: Math Word Problems

Automatically solve math word problems

Question: Tom has **2** apples and Jerry has **5** apples. How many apples do they have in **total**?



Rationale: $x = 2 + 5$

Solution: 7

Four basic arithmetic operations with single or multiple operation steps.



Question: Sara picked 45 pears and Sally picked 11 pears from the pear tree. How many pears were picked in total?

Math ability: basic math

Language complexity: simple language

Format: generative question answering

Knowledge: no external knowledge

Instruction: You are given a question that involves the **calculation of numbers**. You need to perform either an **addition** or **subtraction** operation on the numbers. **Generate your answer** to the given question.

Program 1:

```
def solution(x, y):  
    answer = x + y  
    return answer  
print(solution(45, 11)) # total pears is the sum of  
pears with Sara and Sally
```

Program 2:

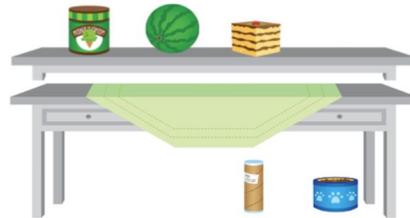
```
x = 45  
y = 11  
answer = x + y # total pears is the sum of pears with  
Sara and Sally  
print(answer)
```

Answer: 56

Multimodal Math Word Problems

Math Reasoning over **Multimodal Information** (images, figures, tables, etc.)

Visual contexts



Q: Which object is next to the one shaped like a cube?



Q: Which picture shows the pizza inside the oven?

C: (A) left one (B) right one



Q: How many sticks are there?

A: 80

IconQA

Tabular contexts



square beads	\$2.97 per kilogram
oval beads	\$3.41 per kilogram
flower-shaped beads	\$2.18 per kilogram
star-shaped beads	\$1.95 per kilogram
heart-shaped beads	\$1.52 per kilogram
spherical beads	\$3.42 per kilogram
rectangular beads	\$1.97 per kilogram

Question: If Tracy buys 5 kilograms of spherical beads, 4 kilograms of star-shaped beads, and 3 kilograms of flower-shaped beads, how much will she spend? (unit: \$)

Answer: 31.44

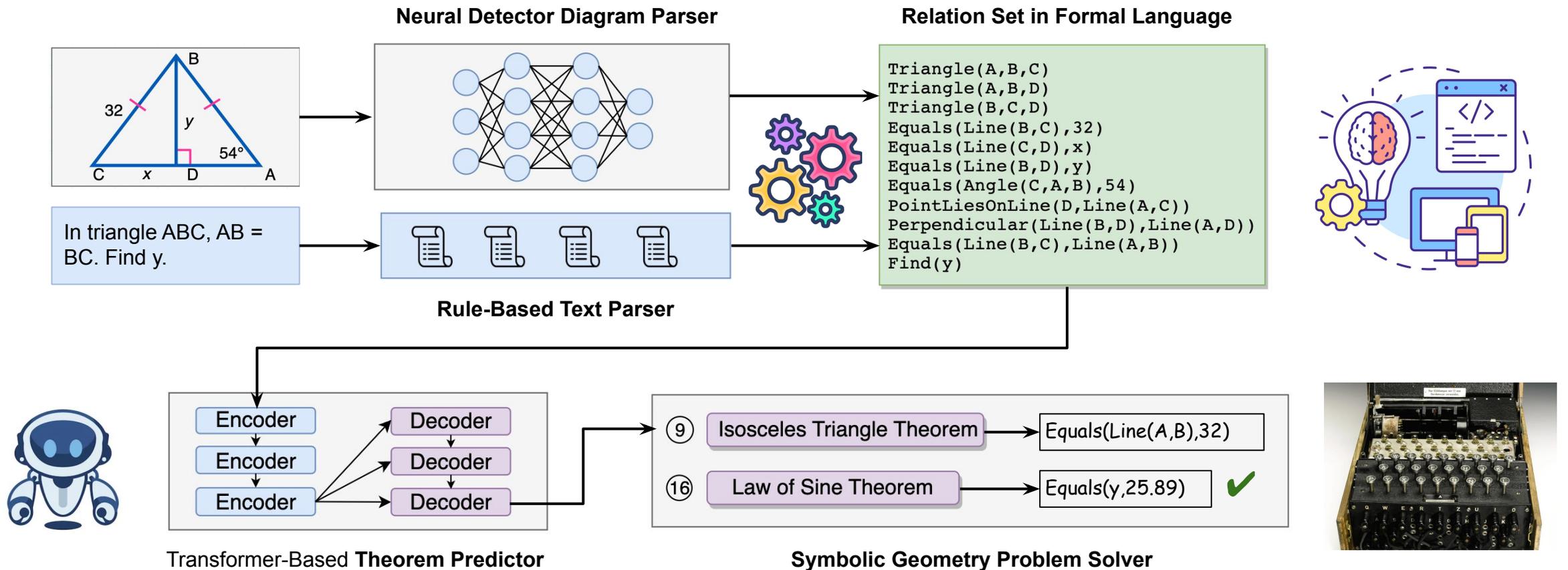
Solution:

Find the cost of the spherical beads. Multiply: $\$3.42 \times 5 = \17.10 .
 Find the cost of the star-shaped beads. Multiply: $\$1.95 \times 4 = \7.80 .
 Find the cost of the flower-shaped beads. Multiply: $\$2.18 \times 3 = \6.54 .
 Now find the total cost by adding: $\$17.10 + \$7.80 + \$6.54 = \31.44 .
 She will spend **\$31.44**.

TabMWP

GPS: Geometry Problem Solving

Neuro-symbolic reasoning over geometry diagrams, theorems, and solvers



Automated Theorem Proving

Demonstrate the truth of a mathematical claim (a **theorem**) via a sequence of logical arguments (a **proof**)

Input: theorem

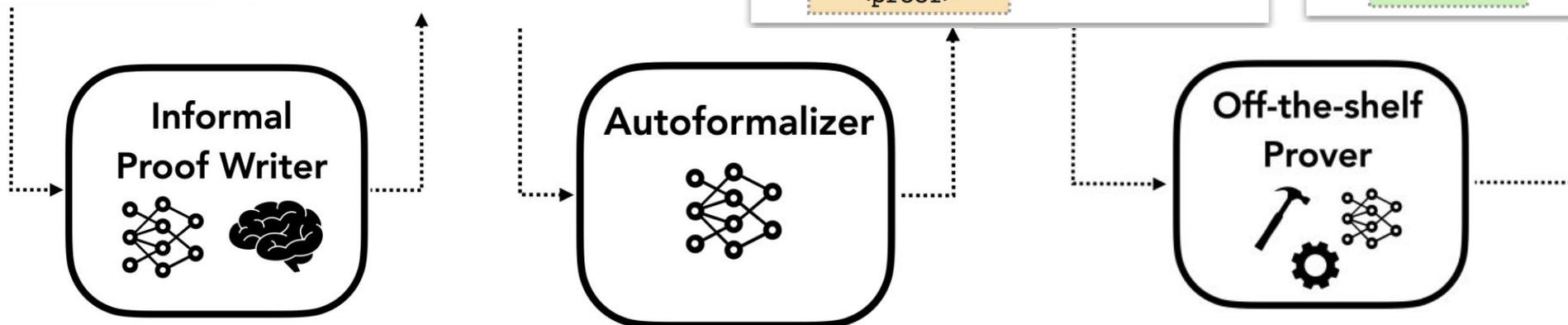
Statement
If $\text{gcd}(n, 4) = 1$ and
 $\text{lcm}(n, 4) = 28$,
show that n is 7.

Informal proof
We know that $\text{gcd}(a, b) \cdot \text{lcm}(a, b) = ab$,
hence $1 \cdot 28 = n \cdot 4$.
Then $n = 1 \cdot 28 / 4 = 7$,
completing the proof. ■

Formal sketch
have c1: "1*28 = n*4"
using assms
<proof>
then have c2: "n = 1*28/4"
<proof>
then show ?thesis
<proof>

Output: proof

Verified formal proof
have c1: "1*28 = n*4"
using assms
by (smt (z3) prod_gcd_lcm_nat)
then have c2: "n = 1*28/4"
by auto
then show ?thesis
by auto



DROP: Reading Comprehension with Discrete Reasoning

Discrete reasoning over content of paragraphs

Passage

Question

Answer

Subtraction



That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.**

How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?

4300000

Comparison



In **1517, the seventeen-year-old King sailed to Castile.** There, his Flemish court **In May 1518, Charles traveled to Barcelona in Aragon.**

Where did Charles travel to first, Castile or Barcelona?

Castile

Selection



In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.

Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?

Don Mueller

Probing Human-level Intelligence of Language Models

Probing Numerical Commonsense Knowledge

Birds can [MASK]. $\xrightarrow[\text{Masked Word Prediction}]{\text{BERT-Large}}$ 1st: fly (79.5%)
2nd: sing (9.1%)

However, for Numerical Commonsense Knowledge :

A bird usually has [MASK] legs.	1st: four (44.8%) 2nd: two (18.7%)
A car usually has [MASK] wheels.	1st: four (53.7%) 2nd: two (20.5%)
A car usually has [MASK] <u>round</u> wheels.	1st: two (37.1%) 2nd: four (20.2%)



BERT

Measuring High-level Problem Solving

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = 7$.

$${}^n C_r = \frac{n!}{r!(n-r)!}$$

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

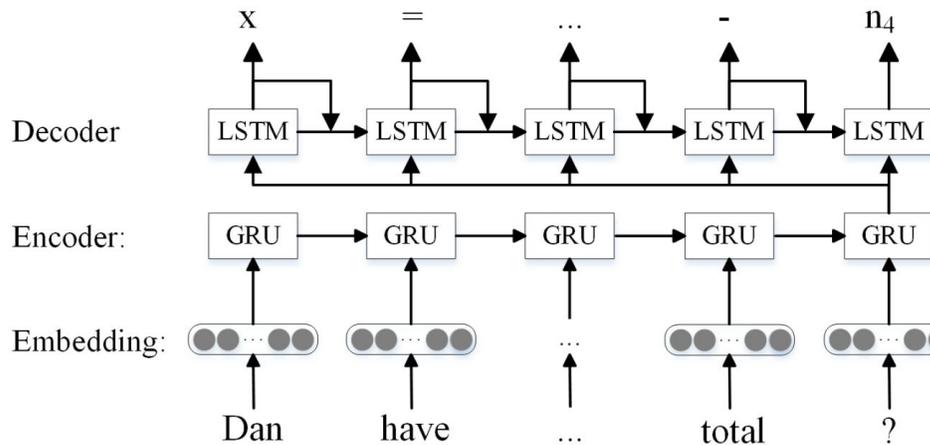
Solution: Complete the square by adding 1 to each side. Then $(x + 1)^2 = 1 + i = e^{\frac{i\pi}{4}} \sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}} \sqrt[4]{2}$. The desired product is then $(-1 + \cos(\frac{\pi}{8}) \sqrt[4]{2})(-1 - \cos(\frac{\pi}{8}) \sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8}) \sqrt{2} = 1 - \frac{(1 + \cos(\frac{\pi}{4}))}{2} \sqrt{2} = \frac{1 - \sqrt{2}}{2}$.

$$i = \sqrt{-1}$$

Seq2Seq Neural Networks

Equation: $x = 5 + 4 + 3 - 2$; solution: [10]

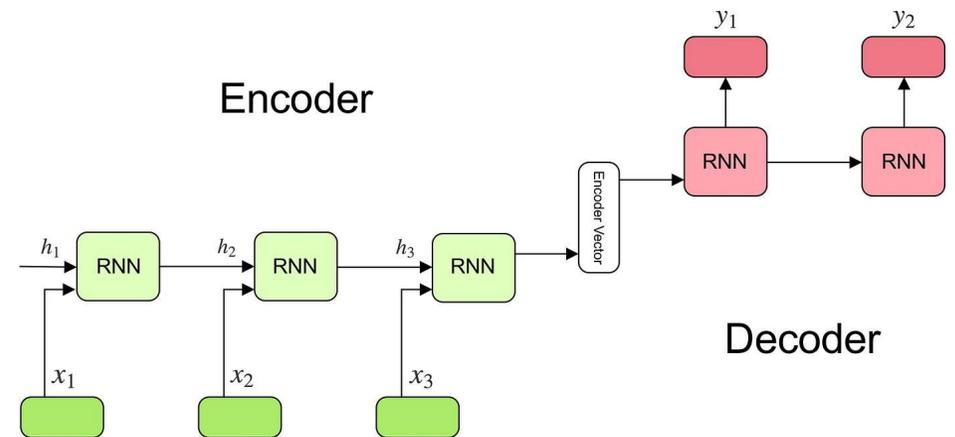
Model output (Equation template): $x = n_1 + n_3 + n_2 - n_4$



Model input: Dan have n_1 , pens and n_2 pencils, Jessica have n_3 more pens and n_4 less pencils than him. How many pens and pencils do Jessica have in total?

Problem: Dan have 5 pens and 3 pencils, Jessica have 4 more pens and 2 less pencils than him. How many pens and pencils do Jessica have in total?

Applying number mapping to equation form



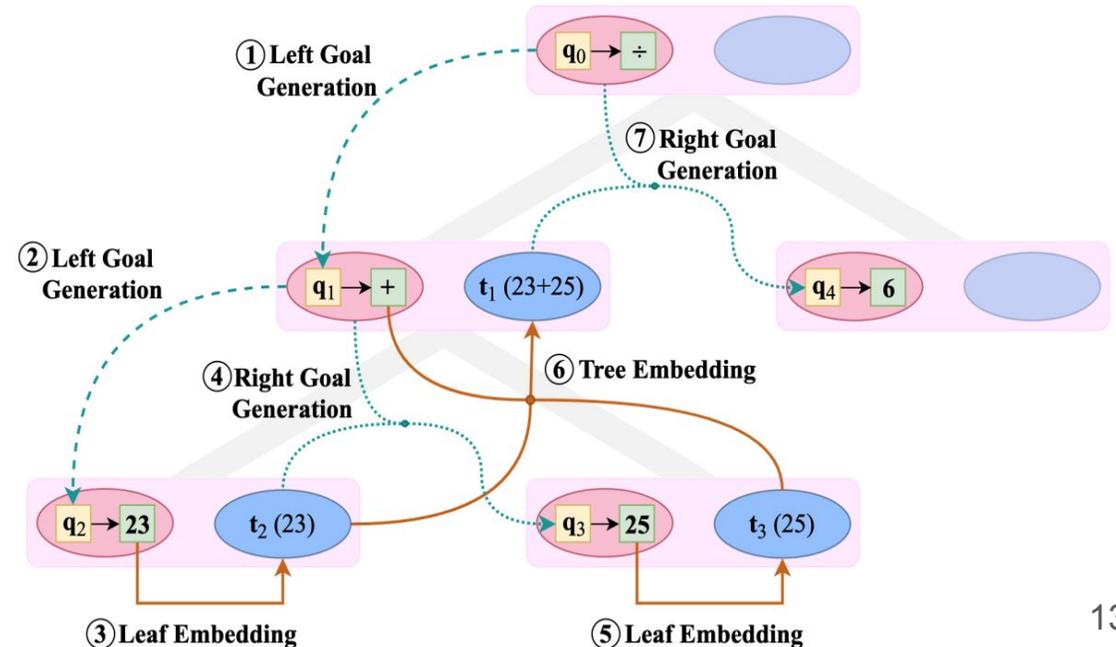
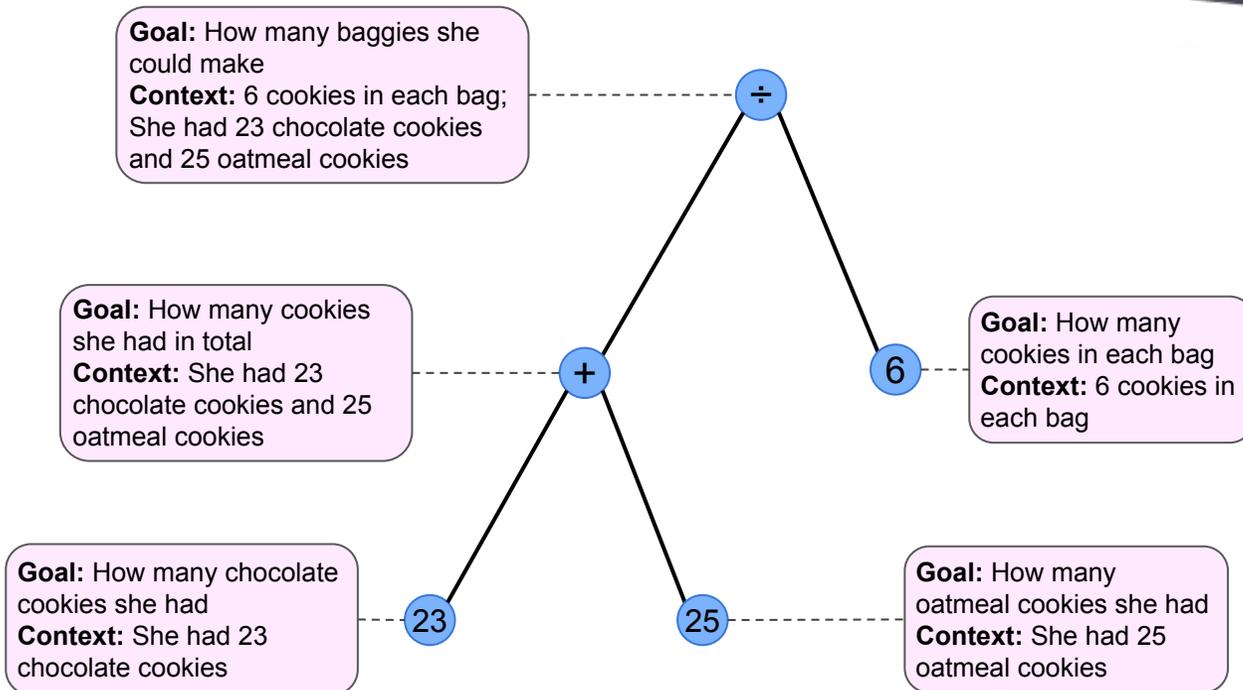
Number mapping:
 $\{n_1=5, n_2=3, n_3=4, n_4=2\}$

Tree-based Neural Networks

Problem: Robin was making baggies of cookies with 6 cookies in each bag. If she had 23 chocolate cookies and 25 oatmeal cookies, how many baggies could she make?

Solution Expression: $(23 + 25) \div 6$
Solution: 8

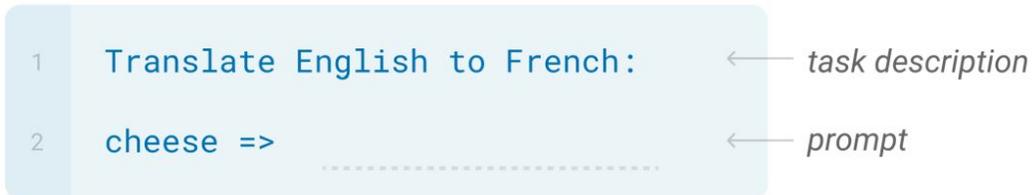
A Goal-Driven Tree-Structured Neural Model for Math Word Problems, IJCAI 2019



LLMs and In-context Learning

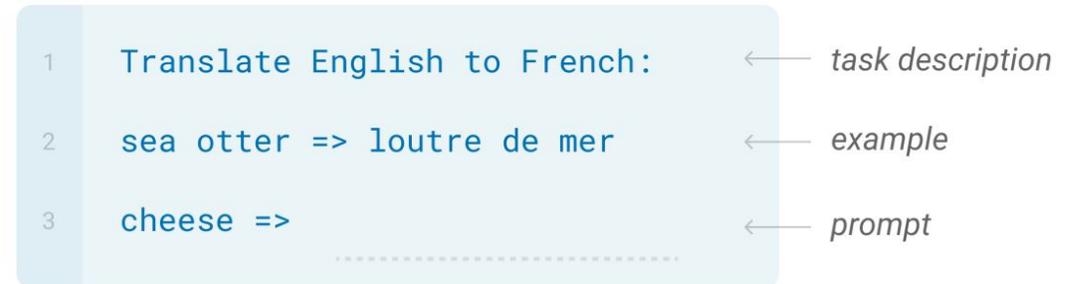
Zero-shot

Given only a natural language **description** of the task



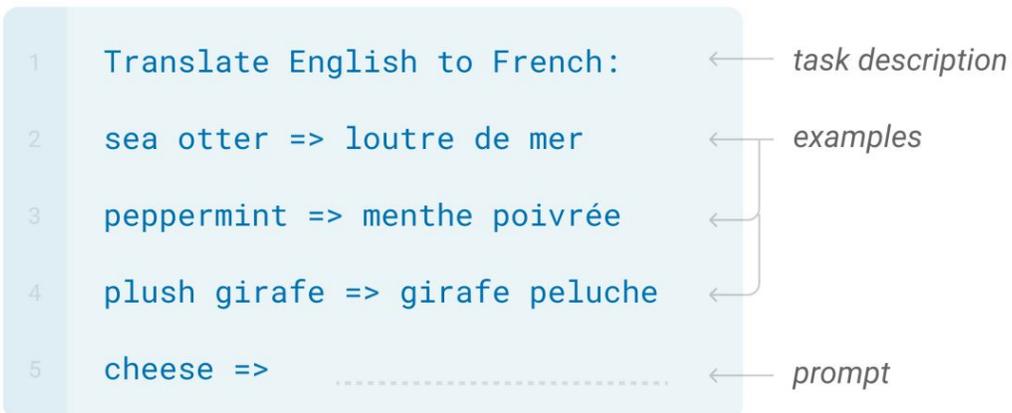
One-shot

Given the task **description** and **a single example** of the task



Few-shot

Given the task **description** and **a few examples** of the task



Emergent Abilities of LLMs

Question: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Standard Prompting

Model 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?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model 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?

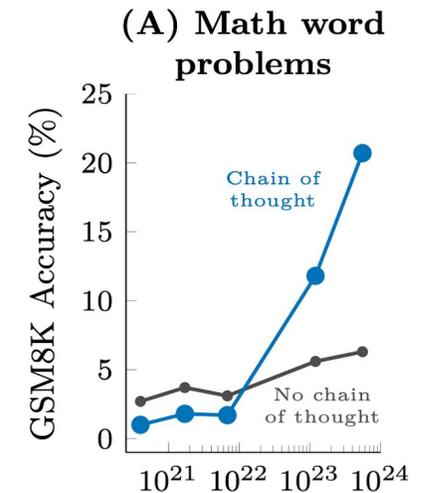
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: 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. ✅

CoT
reasoning
processes



An ability is *emergent* if it is not present in smaller models but is present in larger models.

Chain-of-Thought Prompting (CoT)

Math Word Problems (free response)

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?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

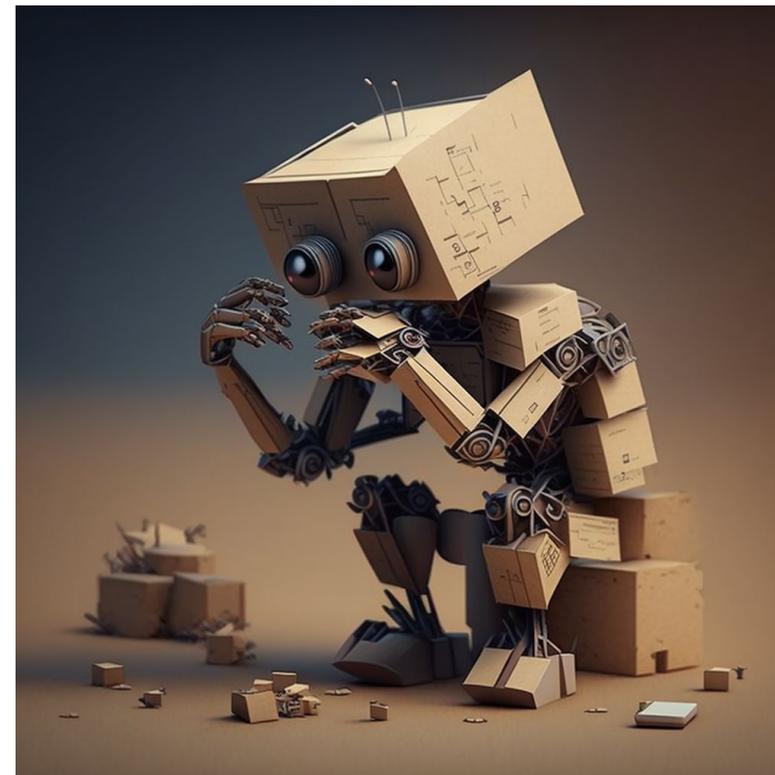
Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.



Chain-of-Thought Prompting (CoT) unlocks and boosts the compositional reasoning ability in the natural language space.

Limitations of LLMs (CoT)

Lack the ability to perform **precise mathematical reasoning**



Table:

Lawn mowers sold	
Day	Number of lawn mowers
Tuesday	8
Wednesday	7
Thursday	5
Friday	7
Saturday	9
Sunday	8
Monday	8

Question: A garden supply store kept track of how many lawn mowers it sold in the past 7 days. What is the median of the numbers?



GPT-3
(text-davinci-002)

Solution:

(Step 1) The median is the "middle" value. To find the median, put the numbers in order from least to greatest and find the number in the middle.

(Step 2) **8, 7, 5, 7, 9, 8, 8**

(Step 3) **The median is 7. The answer is 7.**

Table:

Blocks from school	
Name	Number of blocks
Wendy	14
Carson	8
Zach	17
Lila	17
Nicole	3
Shane	7
Henry	5
Devin	17

Question: Some students compared how many blocks they live from school. What is the mean of the numbers?



GPT-3
(text-davinci-002)

Solution:

(Step 1) The mean is the average of the numbers.

(Step 2) To find the mean, add the numbers and divide by the number of students.

(Step 3) $14 + 8 + 17 + 17 + 3 + 7 + 5 + 17 = 86$

(Step 4) There are 8 students, so divide **86** by 8.

(Step 5) $86 \div 8 = 10.75$

(Step 6) **The mean is 10.75. The answer is 10.75.**

Limitations of LLMs (CoT)

An inability to combine **different tools** to perform complex reasoning



Table:

Tour boat schedule					
Ocean City	8:15 A.M.	9:00 A.M.	9:15 A.M.	9:30 A.M.	10:00 A.M.
Whale Watch Harbor	9:30 A.M.	10:15 A.M.	10:30 A.M.	10:45 A.M.	11:15 A.M.
Oyster Lighthouse	10:15 A.M.	11:00 A.M.	11:15 A.M.	11:30 A.M.	12:00 P.M.
Fisherman's Cove	11:15 A.M.	12:00 P.M.	12:15 P.M.	12:30 P.M.	1:00 P.M.
Surfing Beach	12:00 P.M.	12:45 P.M.	1:00 P.M.	1:15 P.M.	1:45 P.M.

Question: Look at the following schedule. Haley is at Ocean City at 9:45 A.M. How soon can she get to Surfing Beach?

Options: (A) 1:45 P.M. (B) 12:00 P.M. (C) 10:30 A.M. (D) 12:15 P.M.



GPT-3
(text-davinci-002)

Solution:

(Step 1) Find 9:45 A.M. on the schedule.

(Step 2) **The next stop is Whale Watch Harbor at 10:15 A.M.**

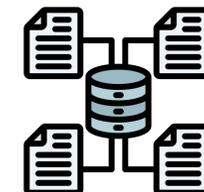
(Step 3) The earliest she can get to Surfing Beach is **12:00 P.M. The answer is 12:00 P.M.**



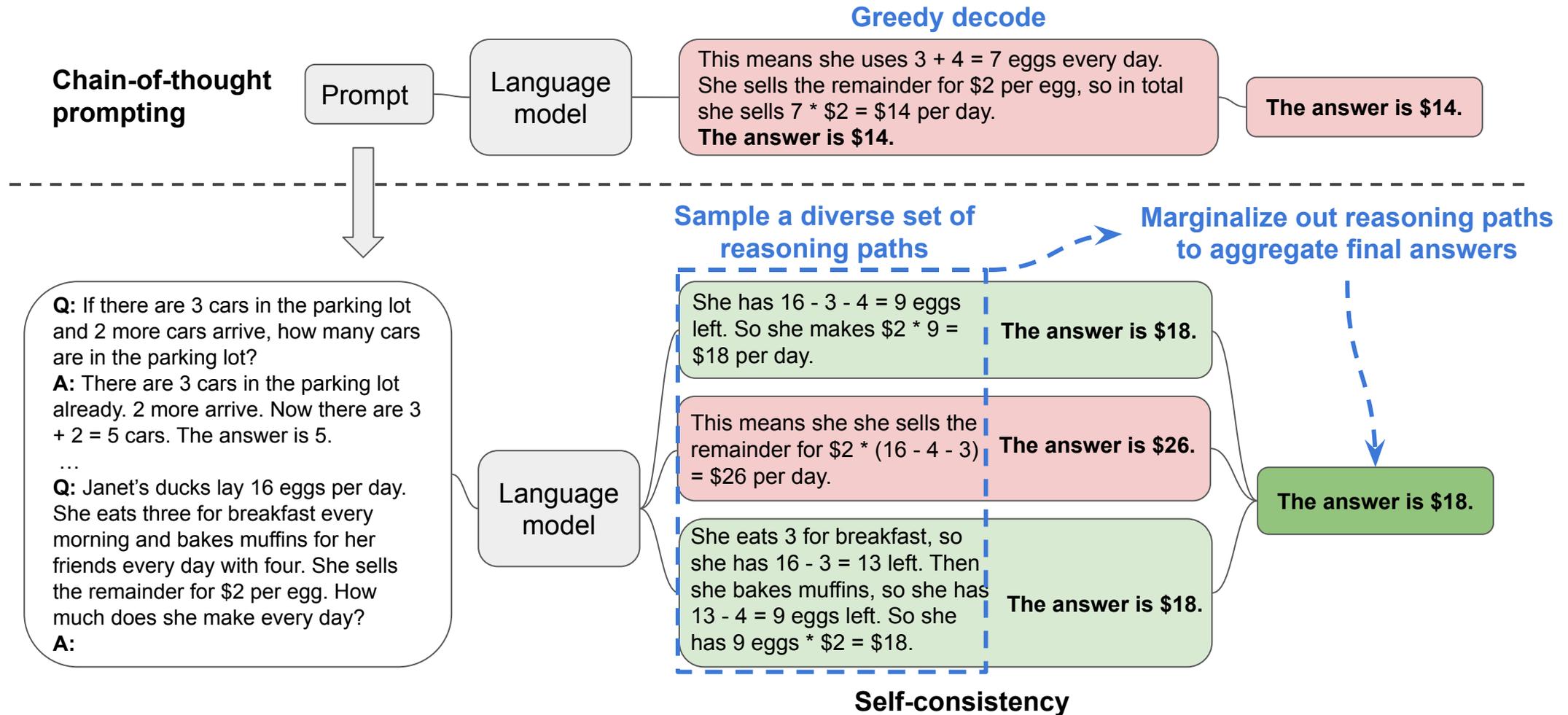
Google

python™

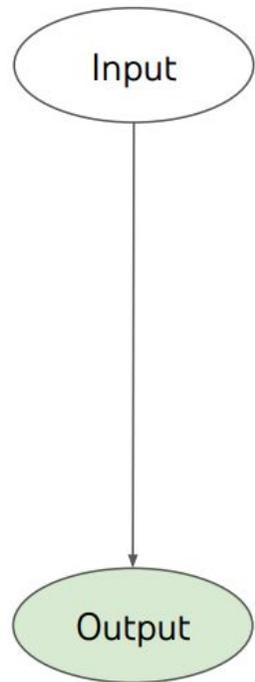
Bing



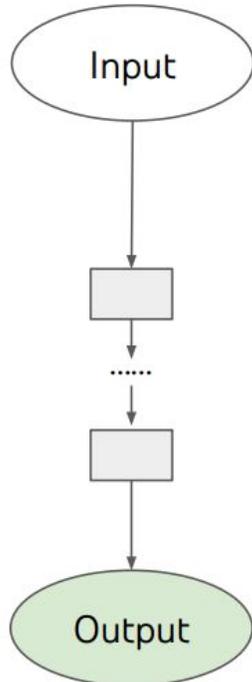
Self Consistency with CoT (CoT-SC)



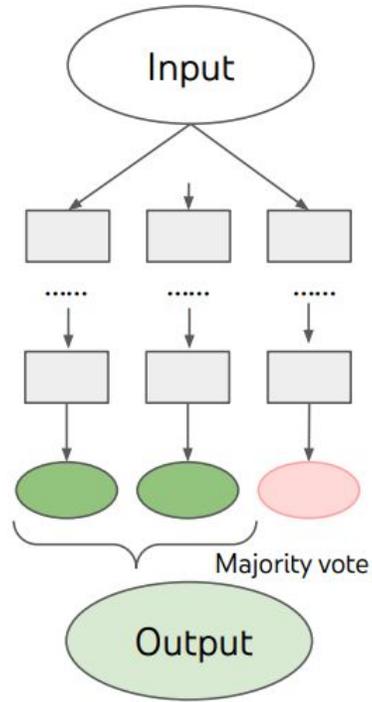
Tree-of-Thought Prompting (ToT)



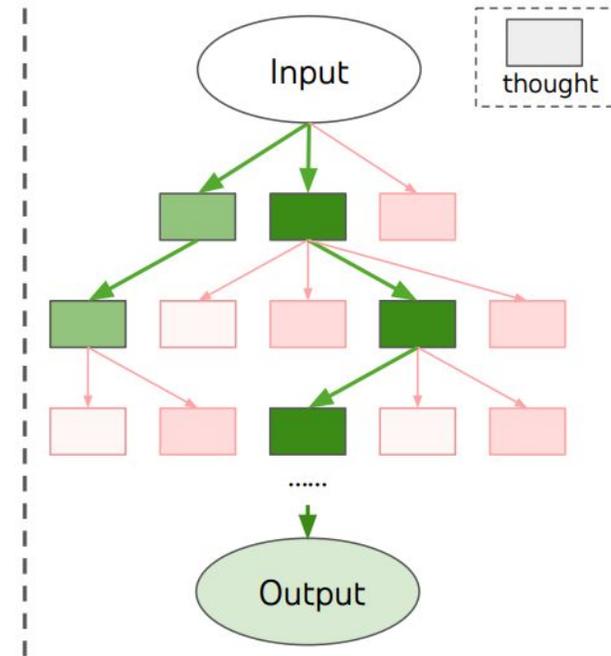
(a) Input-Output Prompting (IO)



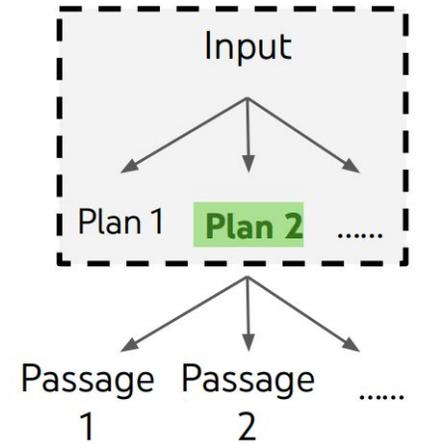
(c) Chain of Thought Prompting (CoT)



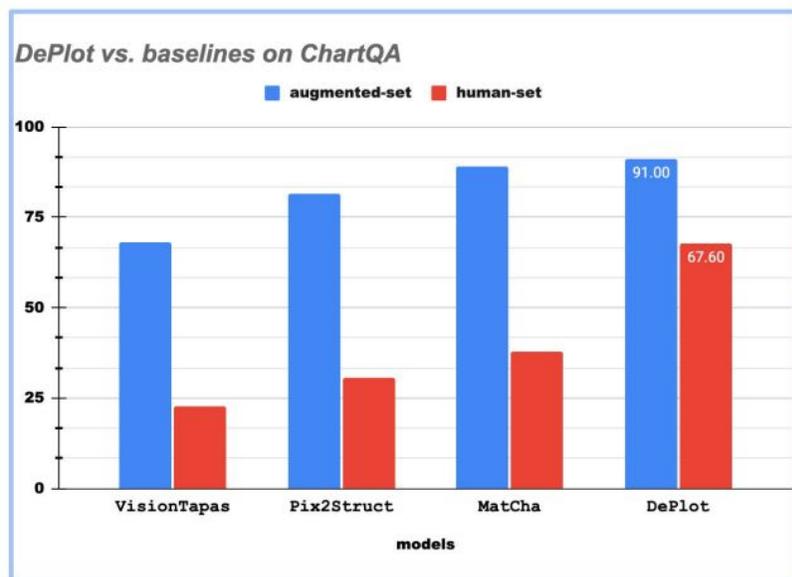
(c) Self Consistency with CoT (CoT-SC)



(d) Tree of Thoughts (ToT)



Tool-Augmented LLMs



DePlot

Header:	models	augmented-set	human-set
Row 1:	VisionTapas	67.2	22.2
Row 2:	Pix2Struct	82.9	30.4
Row 3:	MatCha	89.0	38.0
Row 4:	DePlot	91.0	67.6

LLM

Question: On average across the two sets, by how much does DePlot outperform the second strongest model?

FlanPaLM: Let's find the average of DePlot across the two sets. We add the two numbers in the DePlot row and divide by 2: $91.0+67.6=158.6/2=79.3$. Let's find the second strongest model across the two sets. We find the row with the second highest average: Row 3. We find the average of MatCha across the two sets. We add the two numbers in the MatCha row and divide by 2: $89.0+38.0=127.0/2=63.5$. We subtract the second strongest average from the strongest average: $79.3-63.5=15.8$. The answer is **15.8**.

Program-aided LLMs

Question: In **Fibonacci sequence**, it follows the rule that each number is equal to the sum of the preceding two numbers. Assuming the first two numbers are 0 and 1, what is the 50th number in Fibonacci sequence?

Chain-of-thought (CoT)

The first number is 0, the second number is 1, therefore, the third number is $0+1=1$. The fourth number is $1+1=2$. The fifth number is $1+2=3$. The sixth number is $2+3=5$. The seventh number is $3+5=8$. The eighth number is $5+8=13$.

..... (Skip 1000 tokens)

The 50th number is 32,432,268,459.

⇒ 32,432,268,459



Program-of-thought (PoT)

```
length_of_fibonacci_seq = 50
fibonacci_sequence = np.zeros(length_of_fibonacci_seq)
fibonacci_sequence[0] = 0
fibonacci_sequence[1] = 1
For i in range(3, length_of_fibonacci_seq):
    fibonacci_sequence[i] = fibonacci_sequence[i-1] +
    fibonacci_sequence[i-2]
ans = fibonacci_sequence[-1]
```



⇒ 12,586,269,025



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?

A: Roger started with 5 tennis balls.

```
tennis_balls = 5
```

2 cans of 3 tennis balls each is

```
bought_balls = 2 * 3
```

tennis balls. The answer is

```
answer = tennis_balls + bought_balls
```

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves

```
loaves_baked = 200
```

They sold 93 in the morning and 39 in the afternoon

```
loaves_sold_morning = 93
```

```
loaves_sold_afternoon = 39
```

The grocery store returned 6 loaves.

```
loaves_returned = 6
```

The answer is

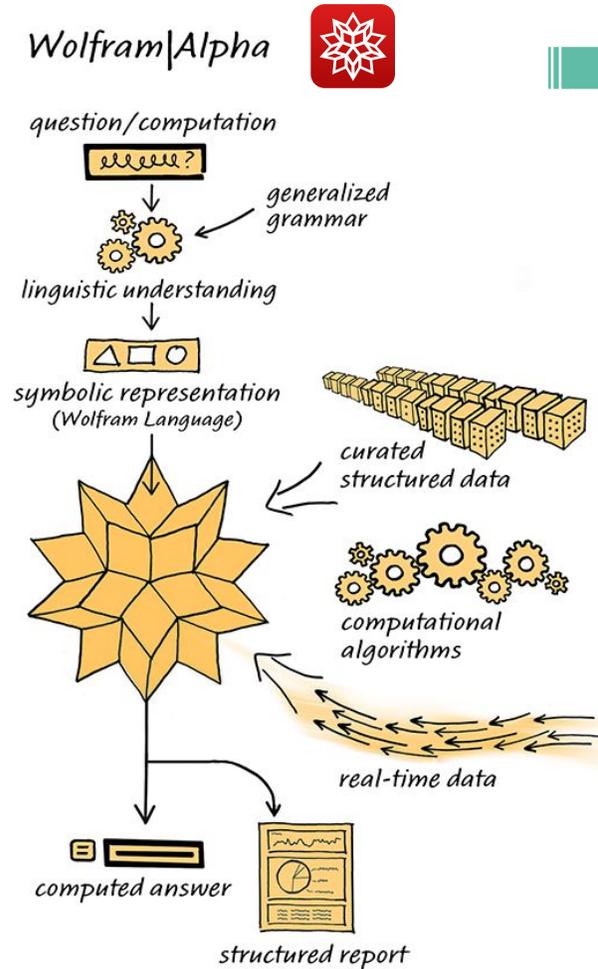
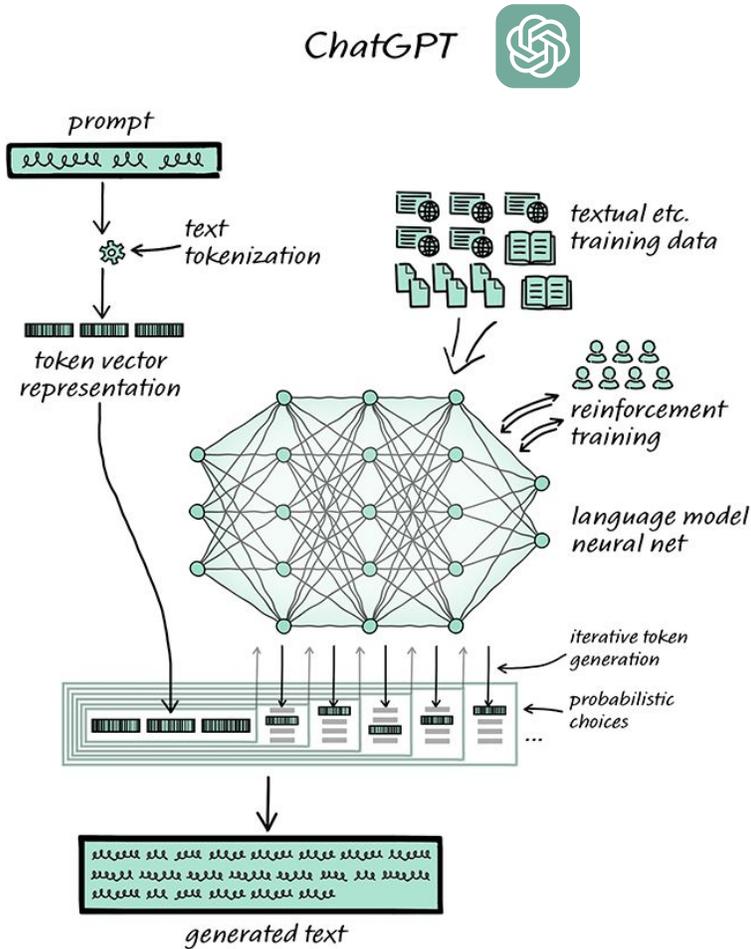
```
answer = loaves_baked - loaves_sold_morning
- loaves_sold_afternoon + loaves_returned
```

```
>>> print(answer)
```

```
74
```



Wolfram Plugin for ChatGPT



Wolfram

Install

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language

S What is the integral of $x^2 \cos(2x)$?

Used **Wolfram**

The integral of $x^2 \cos(2x)$ with respect to x is:

$$\frac{x \cos(2x)}{2} + \frac{(-1 + 2x^2) \sin(2x)}{4} + C$$

where C is the constant of integration.

Chameleon: Plug-and-Play Compositional Reasoning

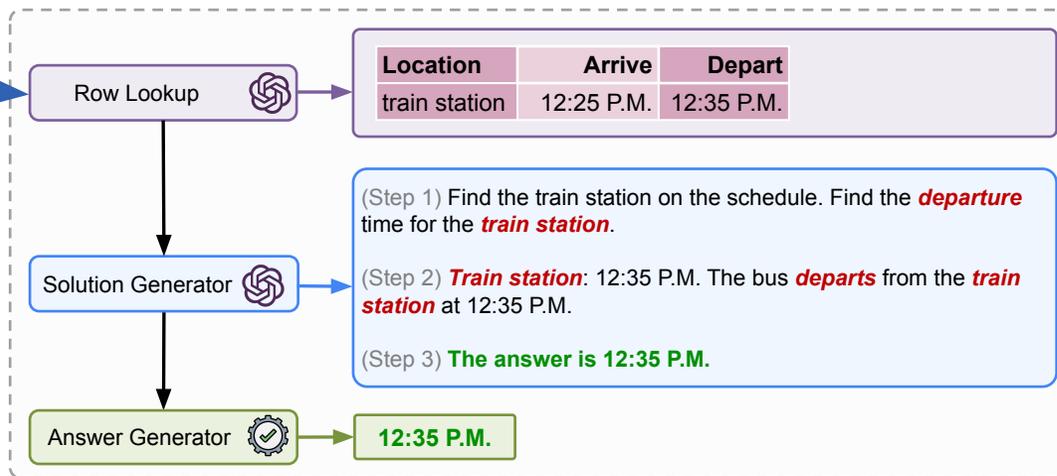
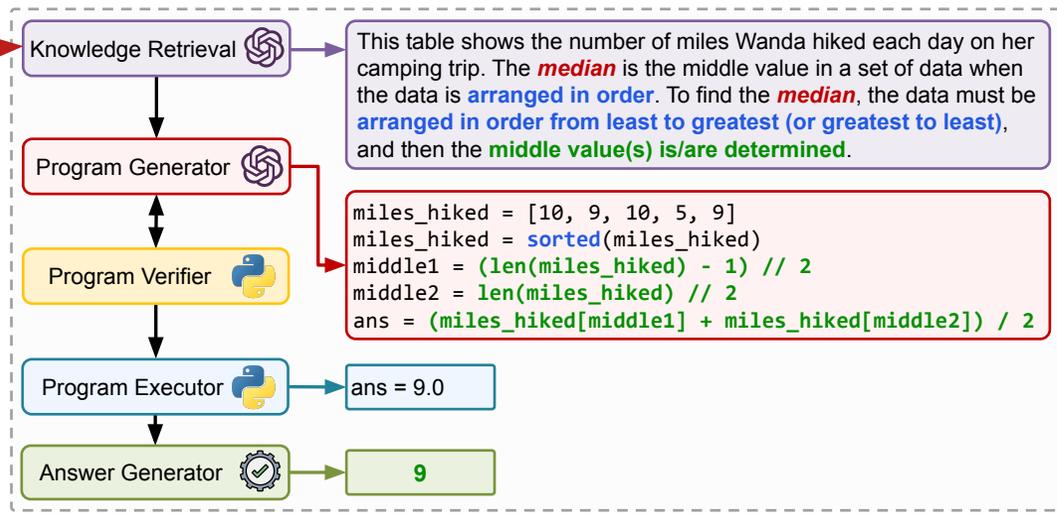
Miles hiked	
Day	Miles
Sunday	10
Monday	9
Tuesday	10
Wednesday	5
Thursday	9

Wanda went on a camping trip and logged the number of miles she hiked each day. What is the **median** of the numbers?

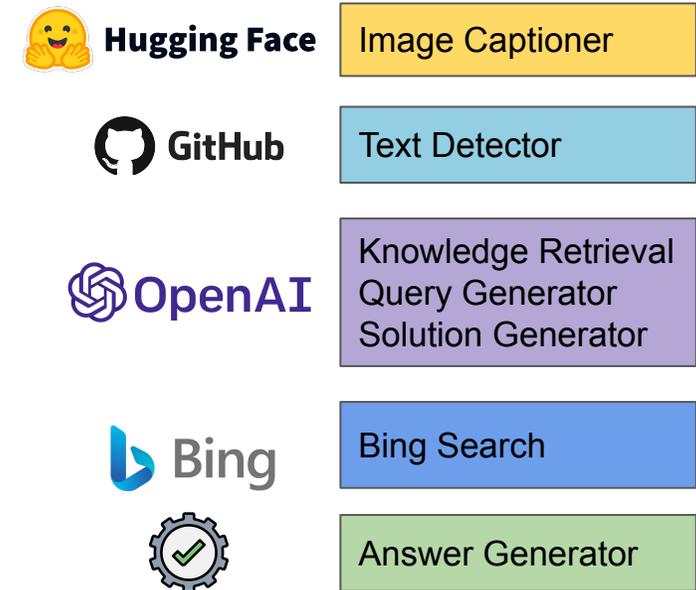
Chameleon

Bus schedule		
Location	Arrive	Depart
stadium	10:20 A.M.	10:25 A.M.
park	10:35 A.M.	10:45 A.M.
hotel	11:10 A.M.	11:15 A.M.
airport	12:05 P.M.	12:10 P.M.
train station	12:25 P.M.	12:35 P.M.
bus station	1:10 P.M.	1:10 P.M.

Look at the following schedule. When does the bus **depart** from the **train station**?
 (A) 12:35 P.M. (B) 1:10 P.M.
 (C) 1:10 P.M. (D) 10:45 A.M.



Tool-augmented LLMs via composing various tools to perform complex tasks.



Low-resource Settings



Annotators check geometric question:

在等腰直角三角形DBC中，角BDC等于90度，BF平分角DBC，与CD相交于点F，延长BD到A，使DA等于DF。求证：三角形FBD与三角形ACD全等。
 In the isosceles right triangle DBC, the angle BDC is equal to 90 degrees. BF bisects the angle DBC and intersects the CD at point F. BD is extended to A making DA is equal to DF. Proof: Triangle FBD and triangle ACD are congruent.

Annotators create:

形状(shape)	等腰直角三角形 (Isosceles Right Triangle)	DBC
	角(Angle)	BDC, DBC
	线段(Line Segment)	BF, CD, BD, DA, DF
	点(Point)	F, A
位置(Position)	交(Intersect), 平分(Bisect)	
数值(Value)	90	
数量关系 (Quantitative Relation)	等于(Equal): (BDC, 90), (DA, DF) 全等(Congruent): (FBD, ACD)	
位置关系 (Positional Relation)	交(Intersect): (BF, CD, F) 平分(Bisect): (BF, DBC)	

Word Problem	Faye had 46 math problems and 9 science problems for homework. If she finished 40 of the problems at school, how many problems did she have to do for homework?
Translation	수정이는 수학 문제 46 문제와 과학 문제 9 문제를 숙제로 풀어야 한다. 수정이가 40 문제를 학교에서 끝냈다면, 더 풀어야 하는 문제는 몇 문제인지 구하여라.
Equation	$X = ((46.0 + 9.0) - 40.0)$
Answer	15.0
Numbers	$n_0 = 46.0, n_1 = 9.0, n_2 = 40.0$
Equation Template	$n_0 n_1 + n_2 -$

Tag	Equation	Segmented	Question
novel	$x = 23 * 2$	اصطاد علي 23 سمكة واصطاد سعود مثليها فما عدد ما اصطاده سعود؟	اصطاد علي 23 سمكة واصطاد سعود مثليها، فما عدد ما اصطاده سعود؟
novel	$x = 18 + 8$	كم عصفورا كان علي الشجرة اذا علمت انه بعد ان طار منهم 8 بقي 18 عصفورا؟	كم عصفورا كان علي الشجرة، إذا علمت أنه بعد أن طار منهم 8 بقي 18 عصفورا؟
inspired	$x = 6 * (2/5)$	كم كيلومتر تساوي (2/5) من 6 كيلومترات =	كم كيلومتر تساوي (2/5) من 6 كيلومترات =
inspired	$x = 160 * 35$	زرع غسان 35 صفا من الازهار في كل صف 160 زهرة؛ فكم زهرة زرع؟	زرع غسان 35 صفاً من الأزهار، في كل صف 160 زهرة؛ فكم زهرة زرع؟
inspired	$x = 50 * 40\%$	ما هو طول القطر للدائرة التي يساوي محيطها 188 سم؟	40% من 50 = .
novel	$x = 188 / 3.14$	أوجد قيمة: 343 - 876 =	ما هو طول القطر للدائرة التي يساوي محيطها 188 سم؟
novel	$x = 876 - 343$	ما هو النظير الضربي ل 8؟	أوجد قيمة: 343-876 =
inspired	$x = 1/8$	ما هو النظير الضربي ل 8؟	ما هو النظير الضربي ل 8؟
inspired	$x = 36/9$	36 شجرة مزروعة في 9 صفوف؛ فكم عدد الأشجار في كل صف؟	36 شجرة مزروعة في 9 صفوف؛ فكم عدد الأشجار في كل صف؟
novel	$x = 677 - 563$	أوجد ناتج: 563 - 677 =	أوجد ناتج: 563-677 =

Low-resource Settings

Finance



Document:

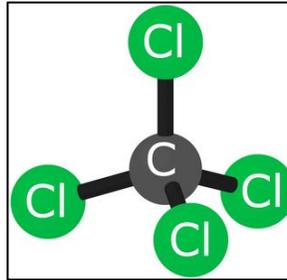
(... abbreviate...)

The following table presents product and service sales and operating expenses by segment (dollar in millions):

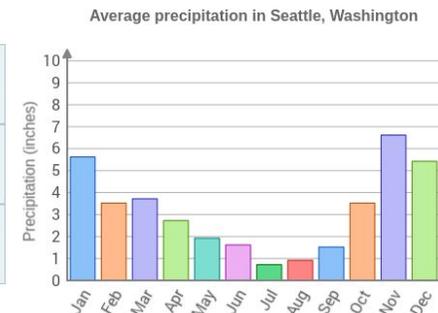
hierarchical row headers		hierarchical column headers			
		Year Ended December 31			
		2018		2017	
Segment		Sales	Expenses	Sales	Expenses
Innovation Systems					
Product		2,894	2,582	—	—
Service		382	351	—	—
Aerospace Systems					
Product		11,087	9,889	10,064	8,988
Service		2,009	1,796	2,067	1,854
Mission Systems					
Product		7,329	6,335	7,012	6,088
Service		4,380	3,854	4,458	3,940
Technology Service					
Product		485	450	391	360
Service		3,812	3,404	4,296	3,878

Product sales for 2018 increased \$4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of \$2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

Science



Planet	Volume (billions of km ³)	Primary composition
Mercury	60	rock
Venus	930	rock
Earth	1,090	rock
Mars	160	rock
Jupiter	1,431,280	gas
Saturn	827,130	gas
Uranus	68,330	ice
Neptune	62,530	ice

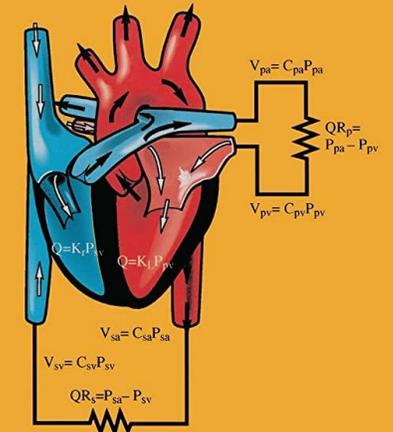


Medicine



F.C. Hoppensteadt C.S. Peskin

Mathematics in Medicine and the Life Sciences

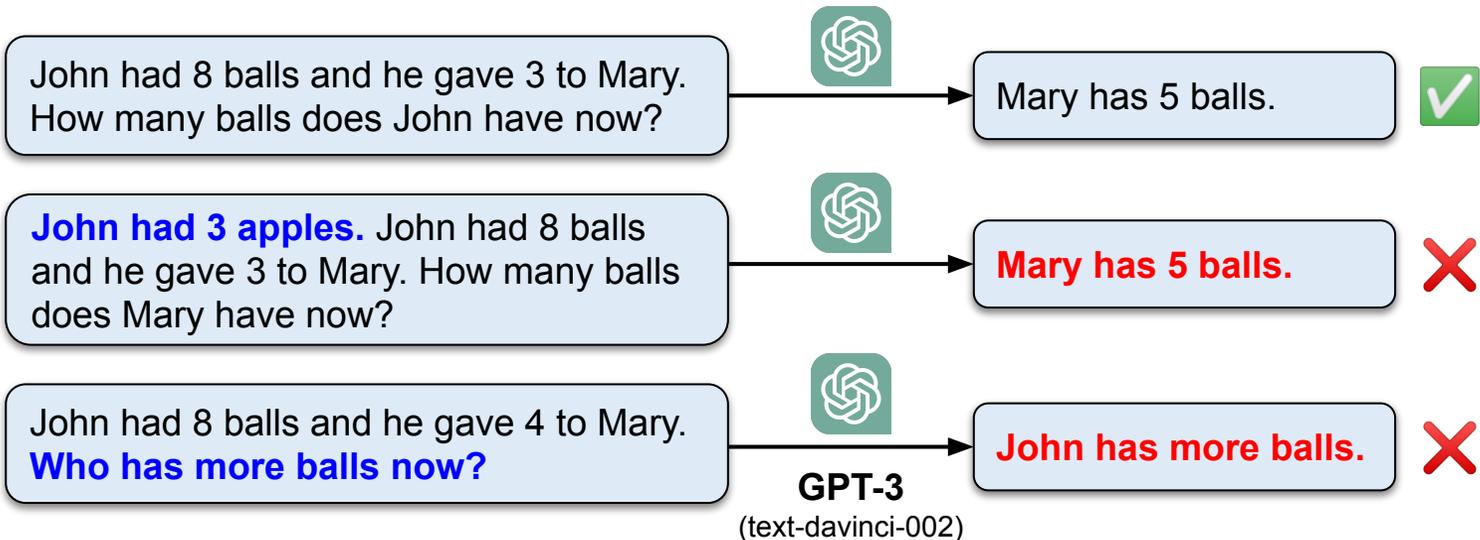


Springer Science+Business Media, LLC

Generalization and Robustness

	T5 (Large)	UnifiedQA (Large)	GPT-3 (text-davinci-002)	GPT-3 (text-davinci-003)
3 balls + 5 balls =	None	5 balls	8 balls	8 balls
23 balls + 145 balls =	None	None	58 balls	168 balls
23 balls + 1,855 balls =	None	None	2,878 balls	2,988 balls

Language models struggle with **large numbers**.

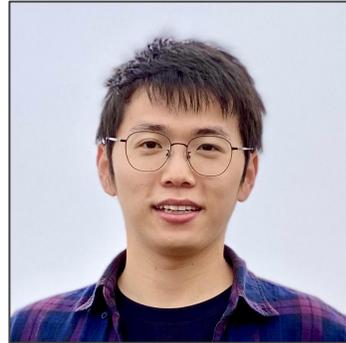


Large language models are **inconsistent** for mathematical reasoning.

IJCAI 2023 Tutorial: Mathematical Reasoning



Zhenwen Liang



Pan Lu



Ashwin Kalyan



Sean Welleck

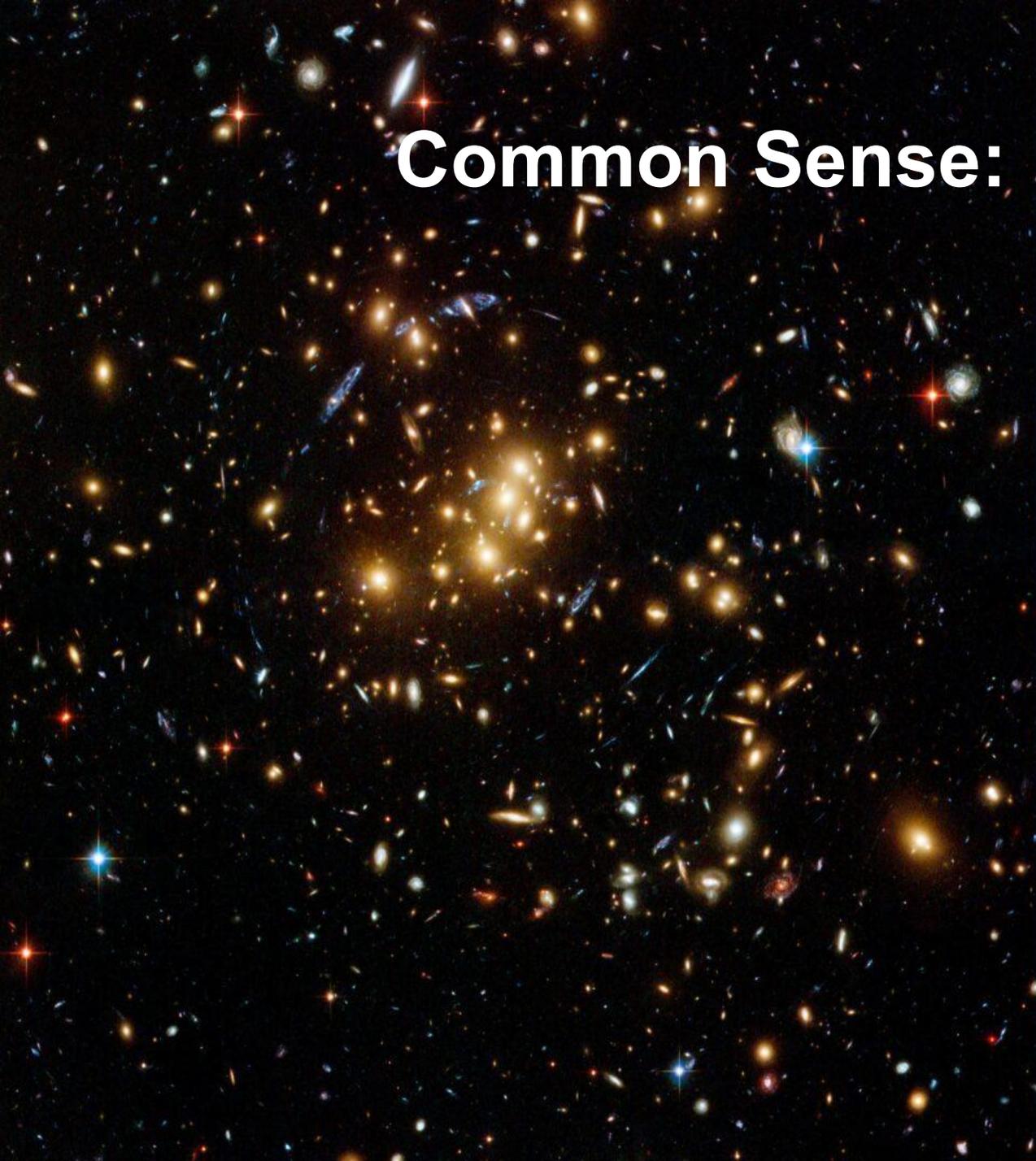


IJCAI/2023 MACAO

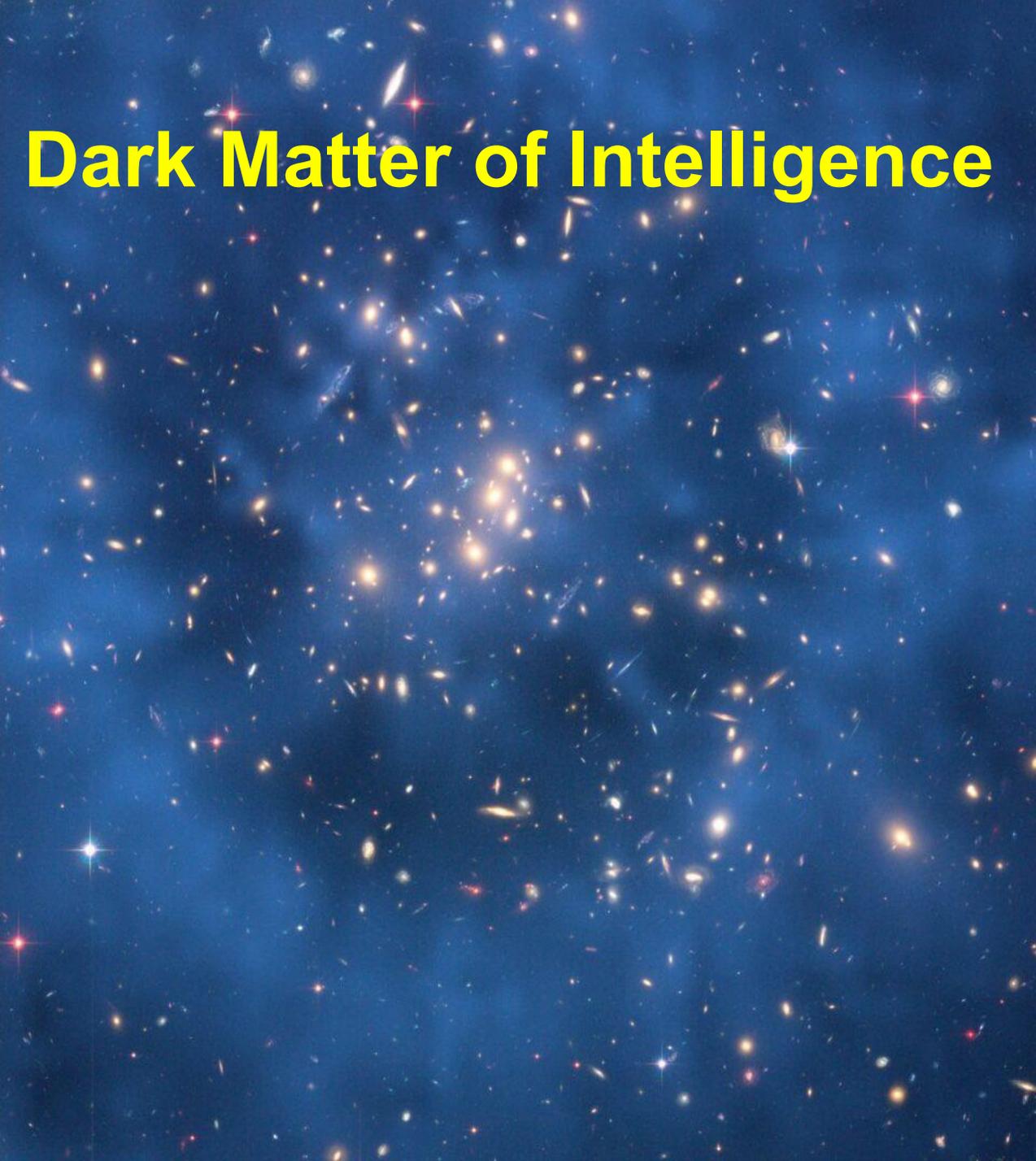
Date: August 21, 2023
Venue: Macao, S.A.R

Registration is open: <https://registration.ijcai.org>



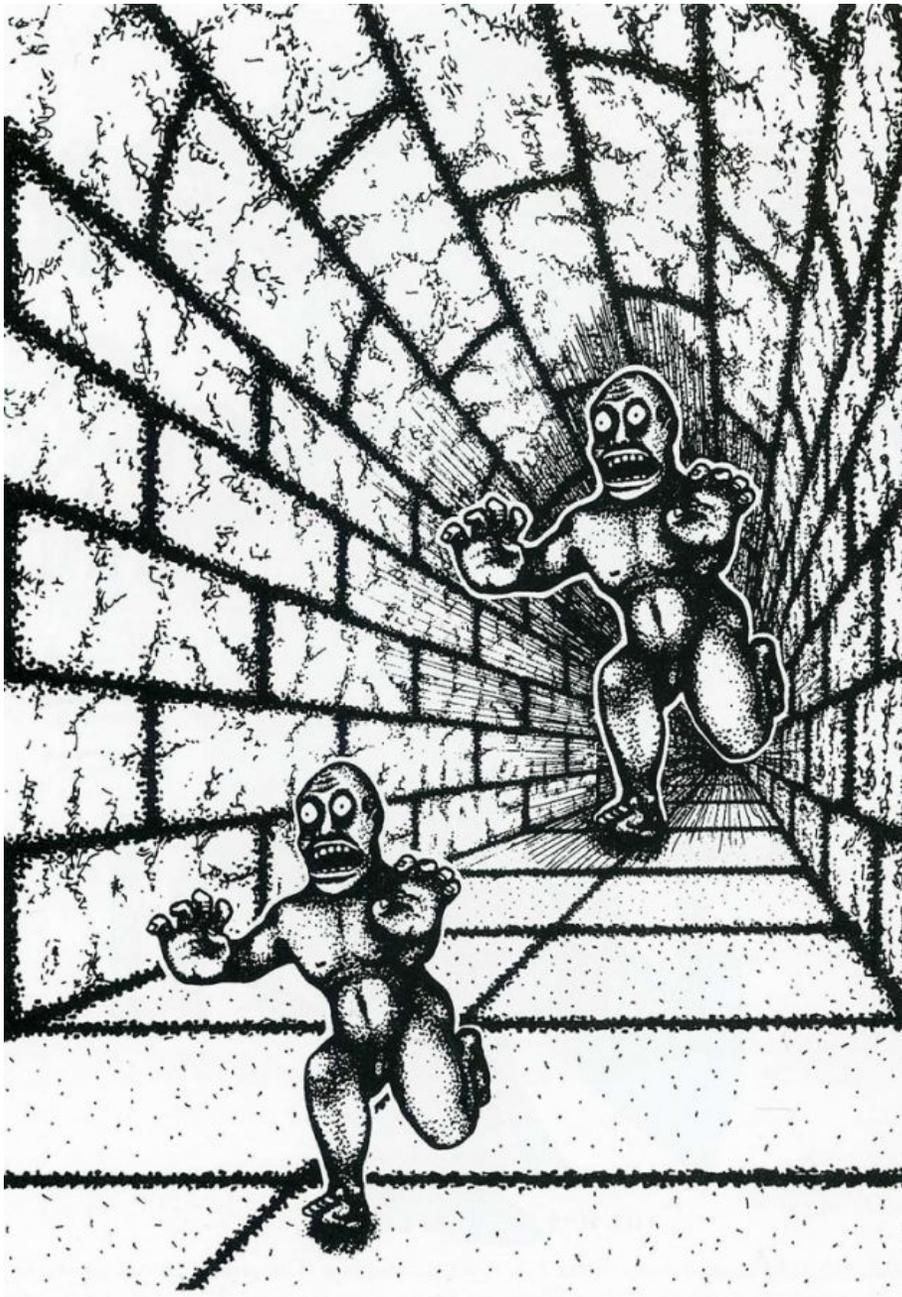


Common Sense:



Dark Matter of Intelligence

Terror Subterra by Roger Shepard



Source: Shepard RN (1990) *Mind Sights: Original Visual Illusions, Ambiguities, and other Anomalies*, New York: WH Freeman and Company



LLaMA-Adapter V2

The figure is a black and white drawing of **two men**, one **standing** and the other **sitting**, with their backs turned to each other. They are positioned in a narrow, dark tunnel or passageway.

<http://llama-adapter.opengvlab.com/>



From human

Two **monsters** in a tunnel (rather than human)

They are **running** (rather than standing or sitting)

One is **chasing** another (rather than the same actions)

The chaser is **hostile** and the chased is **afraid** (even they have the same face expression)



SYSTEM 1
Intuition

95%

Unconscious
Fast
Associative
Automatic pilot



Definition of Common Sense

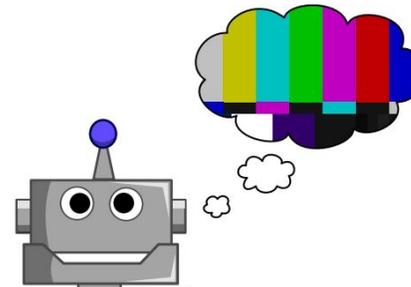
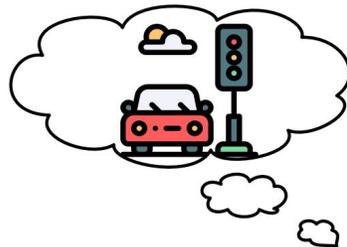
- The basic level of **practical knowledge** and **reasoning**
- Concerning **everyday situations** and **events**
- That are **commonly** shared among **most** people.



For example, it's ok to keep the closet door open, but it's not ok to keep the fridge door open, as the food inside might go bad.



Essential for humans to live and interact with each other in a reasonable and safe way



Essential for AI to understand human needs and actions better

Knowledge in LMs

Do pre-trained LMs **already** have **some** commonsense knowledge?

Birds usually can [MASK].



BERT
(Masked LMs)

Prediction	Score
Birds usually can fly.	33.1% ✓
Birds usually can sing.	8.2% ✓
Birds usually can survive.	3.5% ✓

AllenNLP (MLM [demo](#))



What is the man holding in his hands?



LLaMA-Adapter V2
(Multimodal LMs)

The man is holding a camera in his hands.

Knowledge in LMs

Pre-trained LMs still **lack** commonsense knowledge and can generate **false facts**.

Birds usually can [MASK].



Tigers usually have [MASK] legs.



BERT
(Masked LMs)

Prediction	Score
Birds usually can fly .	<div style="width: 33.1%;"></div> 33.1% ✓
Birds usually can sing .	<div style="width: 8.2%;"></div> 8.2% ✓
Birds usually can survive .	<div style="width: 3.5%;"></div> 3.5% ✓

Prediction	Score
Tigers usually have two legs.	<div style="width: 14.1%;"></div> 14.1% ✗
Tigers usually have short legs.	<div style="width: 11.2%;"></div> 11.2% 🤔
Tigers usually have four legs.	<div style="width: 8.8%;"></div> 8.8% ✓

Commonsense Benchmarks

Social commonsense

Naïve Psychology

ROC story

Social IQa

WSC

COPA



VCR

WinoGrande

Abductive NLI

CommonsenseQA

Physical commonsense

Physical IQa

HellaSwag

SWAG

JHU Ordinal Commonsense



TMTaco

Temporal commonsense



ScienceQA

Scientific commonsense

ReCORD

CosmosQA



MultiRC

Commonsense reading comprehension

Social IQA: Social Interaction QA

Reasoning about **Social Situations**



Kevin spills the chili all over the floor and made the office a huge mess.

What should Kevin do next?

run away and leave the mess in the office

mop up the mess

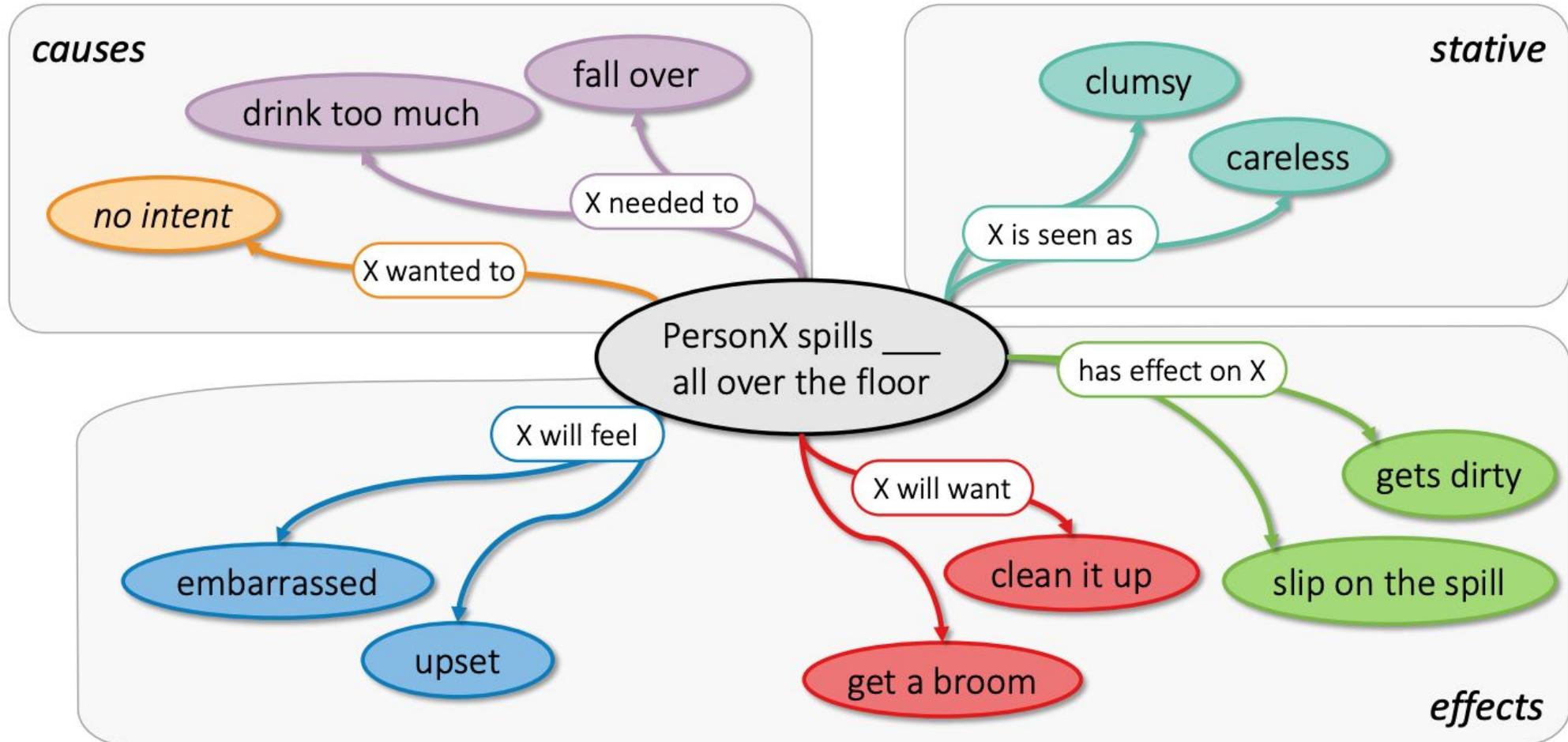
save the chili on the floor



most likely



ATOMIC: Knowledge tested in Social IQA



Physical IQa: Physical Interaction QA

Reasoning about **Physical Interaction**



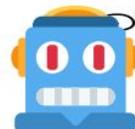
To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.



a!



Test the ability of natural language understanding models to link text to a robust **intuitive-physics** model of the world

ScienceQA: Science Question Answering

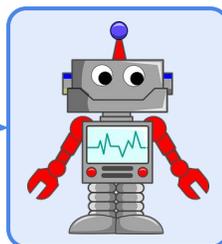
Reasoning about Scientific Topics



Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



Answer: The answer is A.

BECAUSE:



Lecture: A force is a **push** or a **pull** that one object applies to a second object. The direction of a push is **away from** the object that is pushing. The direction of a **pull** is **toward** the object that is pulling.



Explanation: The **baby's hand** applies **a force to** the **cabinet door**. This force causes the **door to open**. The direction of this force is **toward** the **baby's hand**. This force is a **pull**.

ScienceQA: Domain Diversity

<p>Biology</p> <p>Genes to traits Classification Adaptations Traits and heredity Ecosystems Classification Scientific names Heredity Ecological interactions Cells Plants Animals Plant reproduction</p> 	<p>Physics</p> <p>Materials Magnets Velocity and forces Force and motion Particle motion and energy Heat and thermal energy States of matter Kinetic and potential energy Mixture</p> 	<p>Geography</p> <p>State capitals Geography Maps Oceania: geography Physical Geography The Americas: geography Oceans and continents Cities States</p> 	<p>History</p> <p>Colonial America English colonies in North America The American Revolution</p> 	<p>Civics</p> <p>Social skills Government The Constitution</p> 
<p>Earth Science</p> <p>Weather and climate Rocks and minerals Astronomy Fossils Earth events Plate tectonics</p> 	<p>Chemistry</p> <p>Solutions Physical and chemical change Atoms and molecules Chemical reactions</p> 	<p>Writing Strategies</p> <p>Supporting arguments Sentences, fragments, and run-ons Word usage and nuance Creative techniques Audience, purpose, and tone Pronouns and antecedents Persuasive strategies Editing and revising Visual elements Opinion writing</p> 	<p>World History</p> <p>Greece Ancient Mesopotamia World religions American history Medieval Asia</p> 	<p>Economics</p> <p>Basic economic principles Supply and demand Banking and finance</p> 
		<p>Vocabulary</p> <p>Categories Shades of meaning Comprehension strategies Context clues</p> 	<p>Verbs</p> <p>Verb tense</p>	<p>Capitalization</p> <p>Formatting</p>
		<p>Grammar</p> <p>Sentences and fragments Phrases and clauses</p> 	<p>Punctuation</p> <p>Fragments</p>	<p>Phonology</p> <p>Rhyming</p>
		<p>Figurative Language</p> <p>Literary devices</p> 	<p>Reference</p> <p>Research skills</p>	

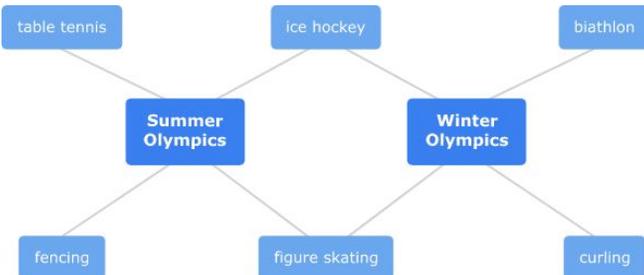
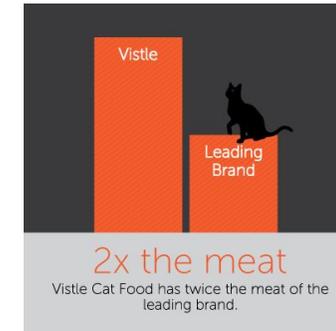
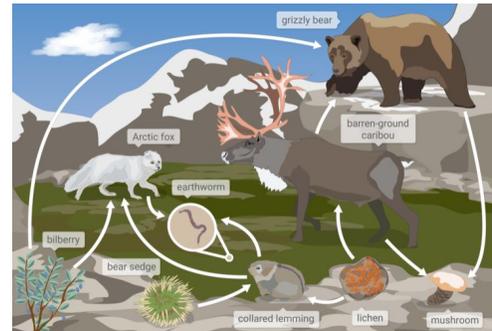
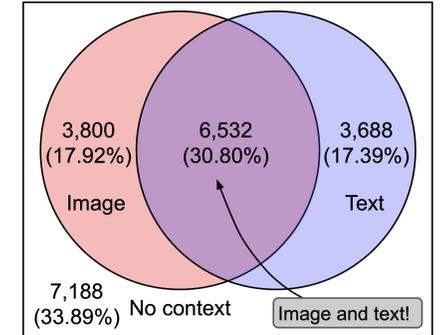
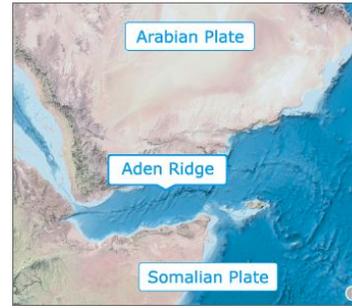
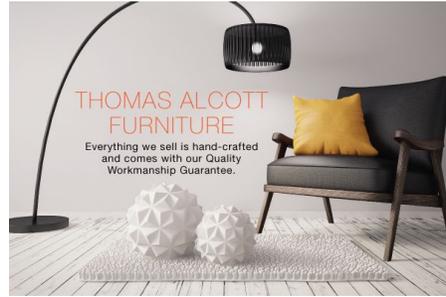
3 subjects

26 topics

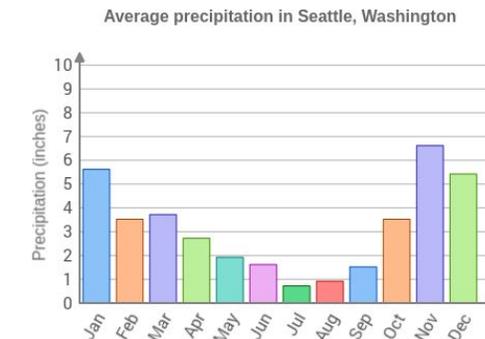
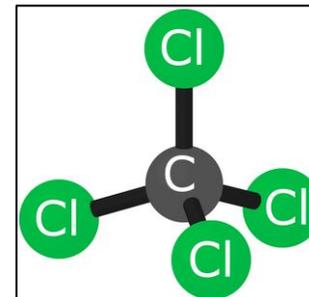
127 categories

379 skills

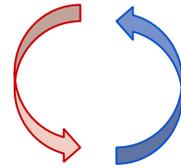
ScienceQA: Context Diversity



Planet	Volume (billions of km ³)	Primary composition
Mercury	60	rock
Venus	930	rock
Earth	1,090	rock
Mars	160	rock
Jupiter	1,431,280	gas
Saturn	827,130	gas
Uranus	68,330	ice
Neptune	62,530	ice



Understanding Time is Important



People were angry

Police used tear gas



People were angry at something (which ended in violent conflicts with the police)...The **police** finally used **tear gas** (to restore order).

Understanding Time is Important



Police used tear gas



People were angry



Police **used tear gas**... **People were angry** with the police.

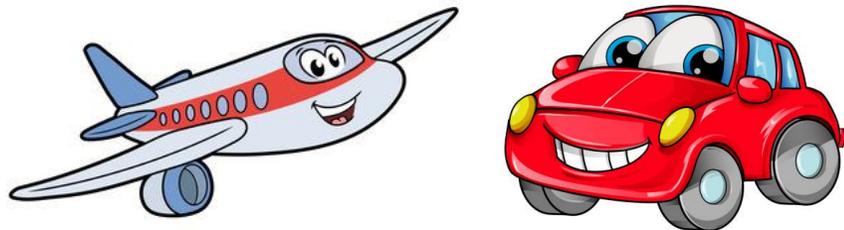
Temporal Commonsense

In natural language, we rarely see explicit **timestamps**.
The language models have to infer the temporal order **from cues in the text**.

- “will” or “will not”?

Dr. Porter is **taking a vacation** and will not be able to see you soon.

Days



Dr. Porter is **taking a walk** and will be able to see you soon.

Minutes



Temporal Commonsense: Challenges

- **Events** are associated with time
- **Attributes** and **relations** may change over time
- **Knowledge bases** need to be qualified temporally



Senator Obama



President Obama



Former President Obama

MC-TACO: Multiple Choice Temporal Common Sense

Reasoning about Temporal Commonsense



Stationarity

S1: Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.

Q1: Is Mark still on the farm now?

no

yes

Reasoning type: stationarity

Typical Time

S2: The massive ice sheet, called a glacier, caused the features on the land you see today.

Q2: When did the glacier start to impact the land's features?

centuries ago

hours ago

10 years ago

tens of millions of

Reasoning type: event typical time

years ago

Duration

S3: Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.

Q3: How long did the tour last?

9 hours

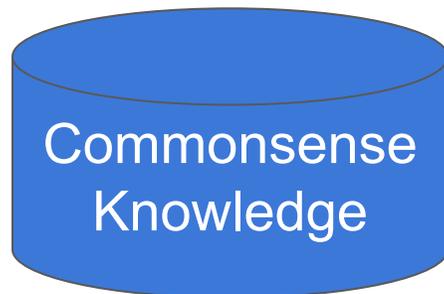
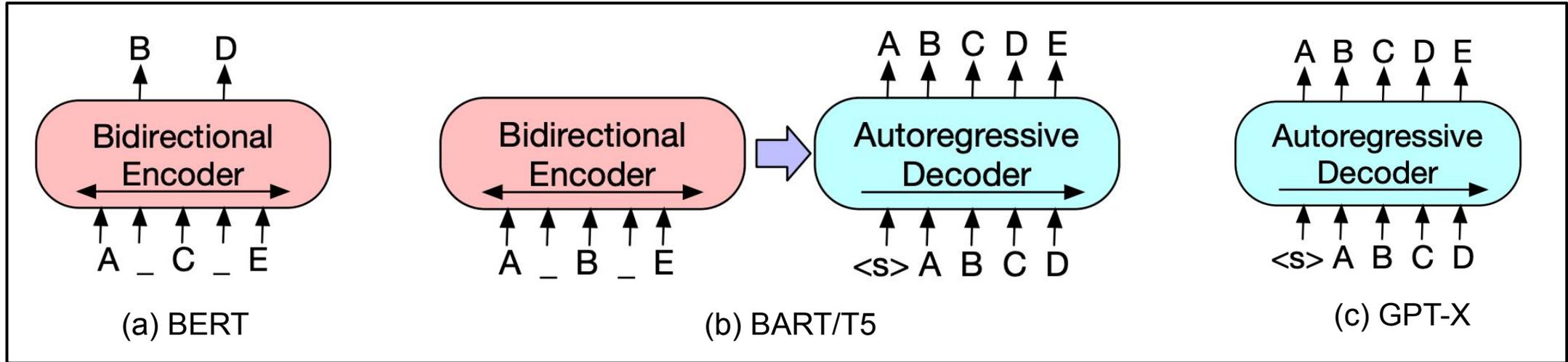
15 days

45 minutes

5 seconds

Reasoning type: event duration

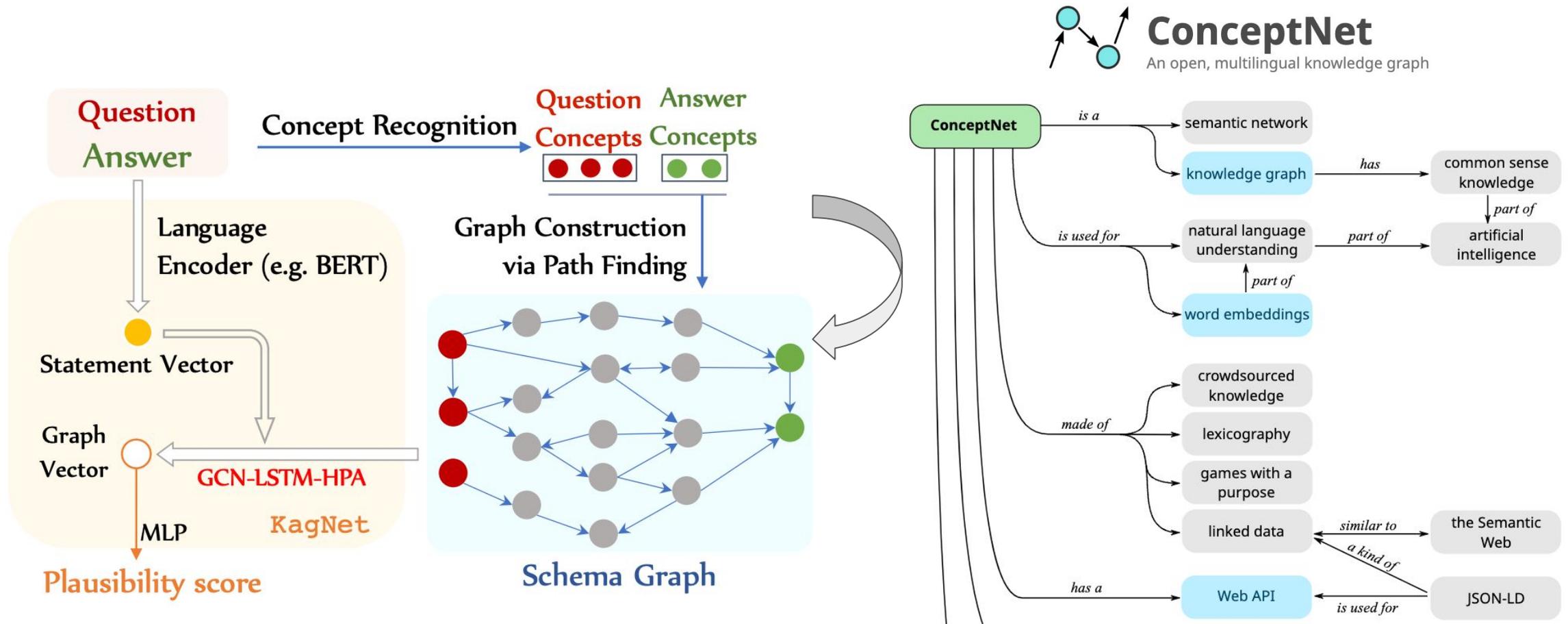
Neural Language Models and Commonsense Knowledge



Higher performance
→
More socially responsible



Incorporating Structured Commonsense Knowledge



Incorporating Unstructured Commonsense Knowledge

Q: What can help alleviate global warming?



Open-Ended CSR
(reason w/ question)

Multiple-Choice CSR
(reason w/ question+choice)
(A) air cooler (B) fossil fuel
(C) **renewable energy** (D) water



a large text corpus of commonsense **facts**

Multi-Hop Reasoning

carbon dioxide is the major **greenhouse gas** contributing to **global warming** .

trees remove **carbon dioxide** from the **atmosphere** through **photosynthesis** .

Output: ..., renewable energy, **tree**, solar battery, ...



a ranked list of concepts as answers.

Open-Ended Commonsense Reasoning



Why AI is incredibly smart -- and shockingly stupid

TED Ideas worth spreading



Why AI is incredibly smart and shockingly stupid

1,051,826 views | Yejin Choi • TED2023

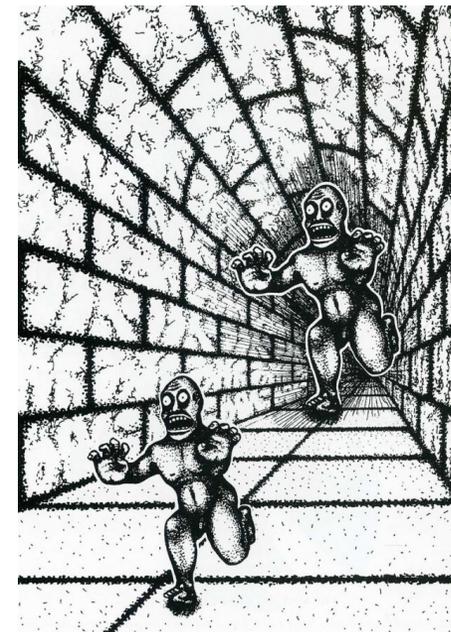
https://www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid/c



Main References & Additional Reading

- [1] Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, Kai-Wei Chang, *A Survey of Deep Learning for Mathematical Reasoning*, ACL 2023.
- [2] *Resources of deep learning for mathematical reasoning (DL4MATH)*: <https://github.com/lupantech/dl4math>.
- [3] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, Dan Roth, *Commonsense Reasoning for Natural Language Processing*, ACL 2020 Tutorial.
- [4] Yuchen Lin, Xiang Ren, *Commonsense Knowledge and Reasoning for NLP*, ACL 2022 KnowledgeNLP Tutorial.
- [5] Yejin Choi, *Why AI is incredibly smart and shockingly stupid*, TED 2023
- [6] Yejin Choi, *The Curious Case of Commonsense Intelligence*, 2022
- [7] Yejin Choi, *Neural Knowledge and Reasoning: On Commonsense Intelligence*, UW CSE P 517 -- Natural Language Processing [Winter 2021].

Thanks for your attention!



These lecture slides are intended solely for educational and research purposes.

The copyright of the content within these slides is held by the respective original creators.

Any utilization of the materials should be accompanied by proper acknowledgment and citation in accordance with relevant copyright and intellectual property guidelines.

Thanks!