Cumulative Reasoning with LLM

Yifan Zhang*, Jingqin Yang*, Yang Yuan, Andrew Chi-Chih Yao

Thinking Fast and Slow

- **Dual modes of thinking**:
  - **System 1 (intuitive and fast) and System 2 (rational and slow).** **--- Daniel Kahneman**

- Currently, LLMs align more closely with **System 1**, thereby potentially explaining their limitations in confronting complex tasks.

- In response to these limitations, several methods have been proposed to mimic human cognitive processes, such as CoT, CoT-SC, ToT, etc.

Chain-of-Thought Prompting

In the pioneering work on **chain-of-thought** reasoning, Wei et al. (2022) emphasized the importance of incorporating multi-step reasoning paths before generating definitive answers.

By few-shot examples, or simply prompting "Let's think step by step".

In a progression from this, Wang et al. (2022b) introduced **self-consistency**, a sophisticated decoding strategy based on majority voting.

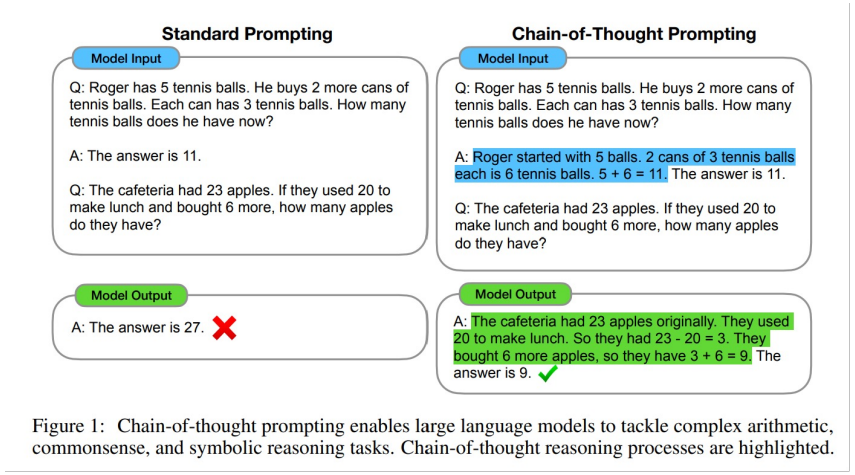


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

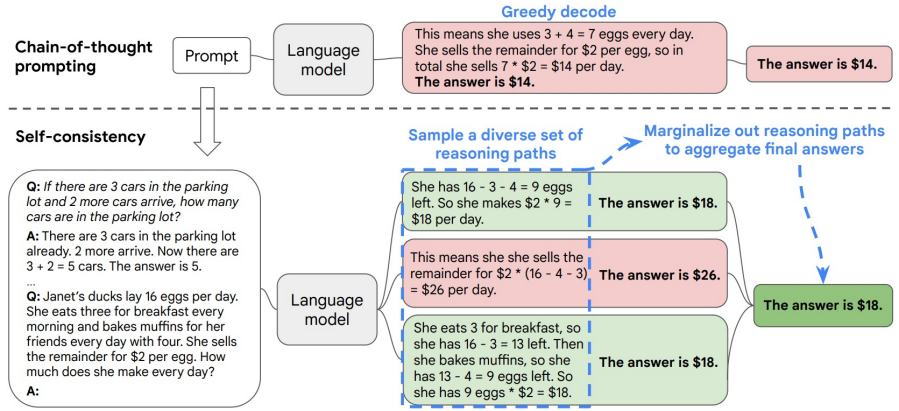


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the "greedy decode" in CoT prompting by sampling from the language model's decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

Tree of Thoughts

Tree-of-Thought (ToT) models the solving process as a thought search tree (Yao et al., 2023; Long, 2023). In addition, dedicated datasets have been created to provide stepwise guidance in model training (Lightman et al., 2023).

Nevertheless, these methods do not have **a site** for storing intermediate results, assuming that all the thoughts form a chain or a tree, which does not fully capture the human thinking process.

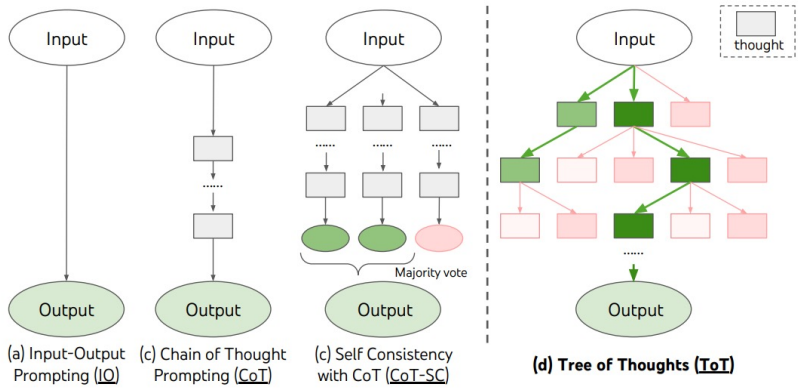


Figure 1: Schematic illustrating various approaches to problem solving with LLMs. Each rectangle box represents a *thought*, which is a coherent language sequence that serves as an intermediate step toward problem solving. See concrete examples of how thoughts are generated, evaluated, and searched in Figures 2,4,6.

Cumulative Reasoning (CR)

Our CR method uses three distinct types of LLMs:

1. Proposer: This model suggests the next step based on the current context.
2. Verifier(s): This model or set of models scrutinizes the accuracy of the step put forward by the proposer. If the step is deemed correct, it will be added to the context.
3. Reporter: This model determines when the reasoning process should be concluded, by accessing whether the current conditions can directly lead to the final solution.

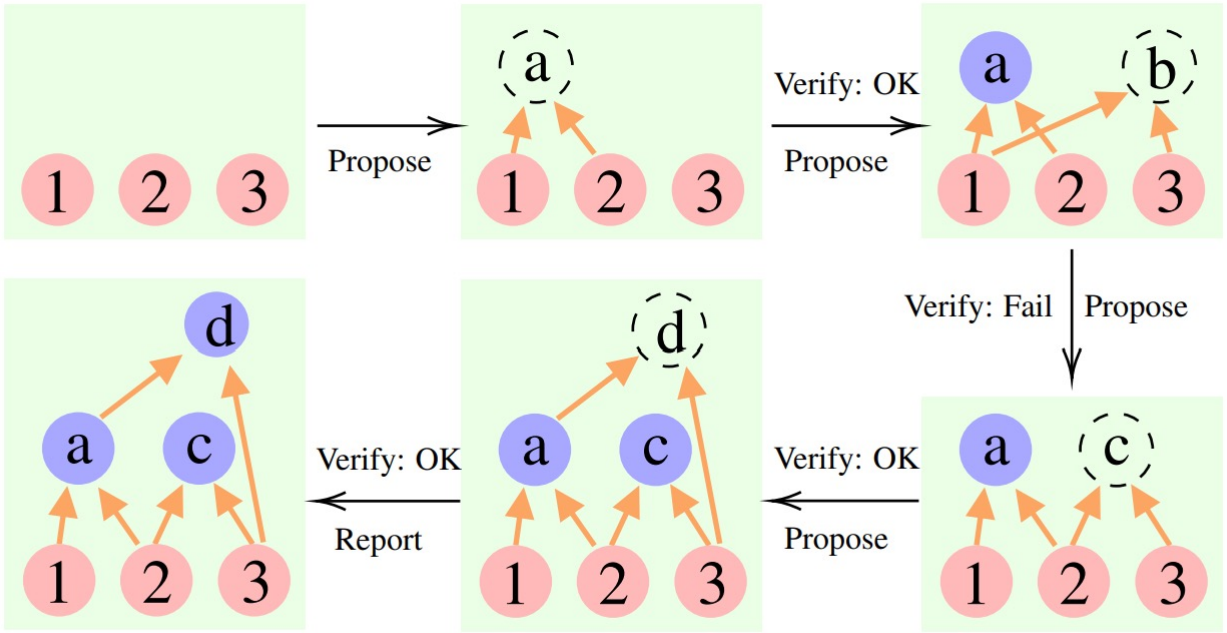
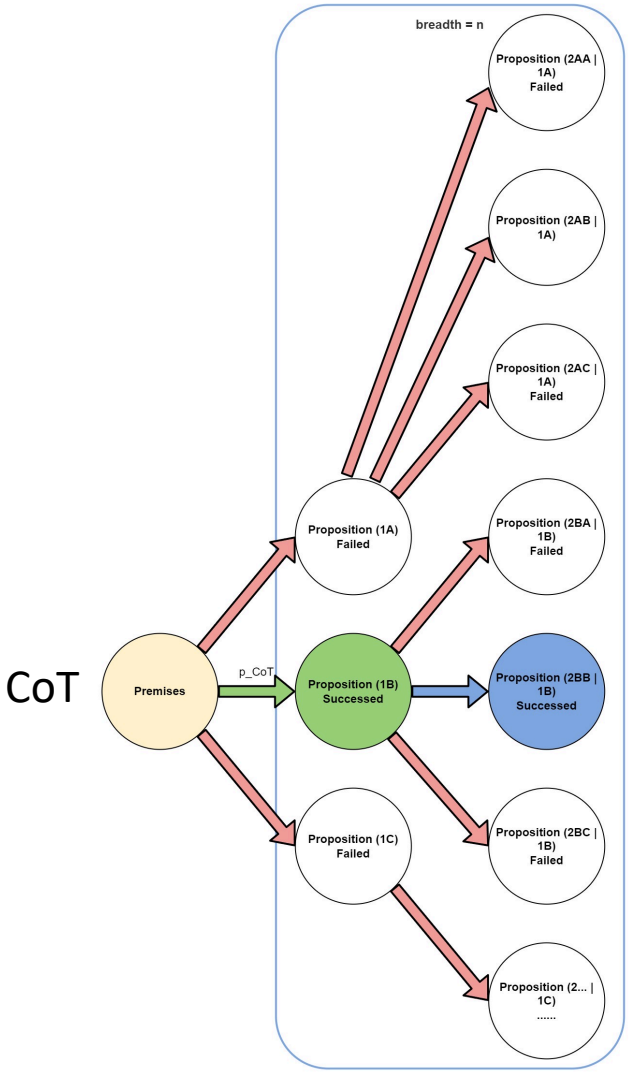
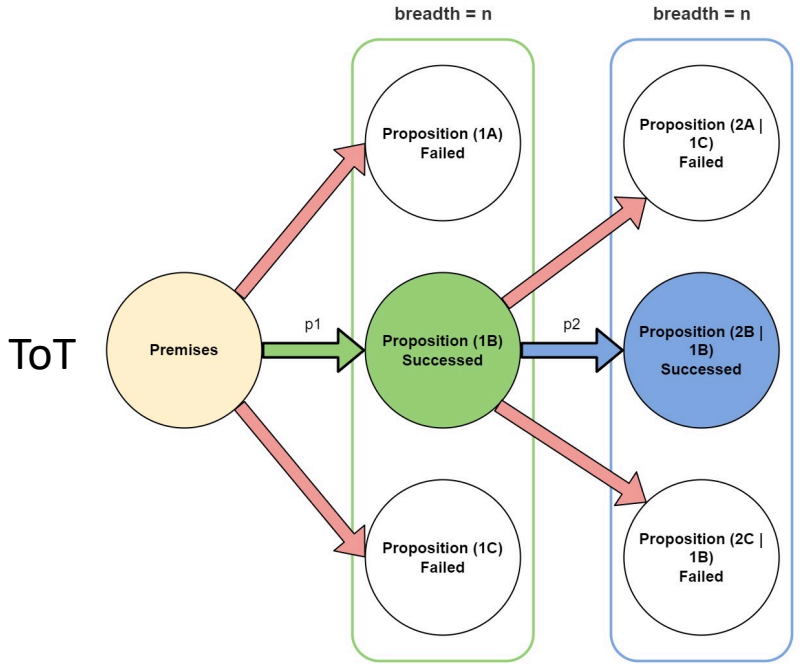


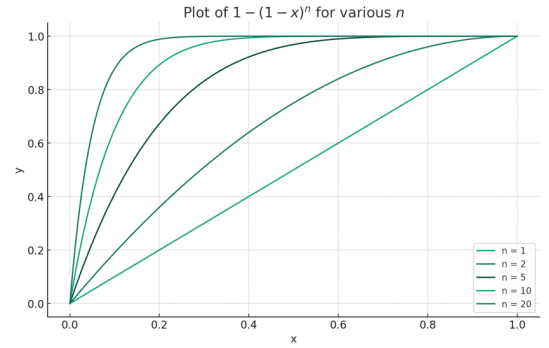
Figure 2: An illustration of CR Reasoning for a 3-premises problem.

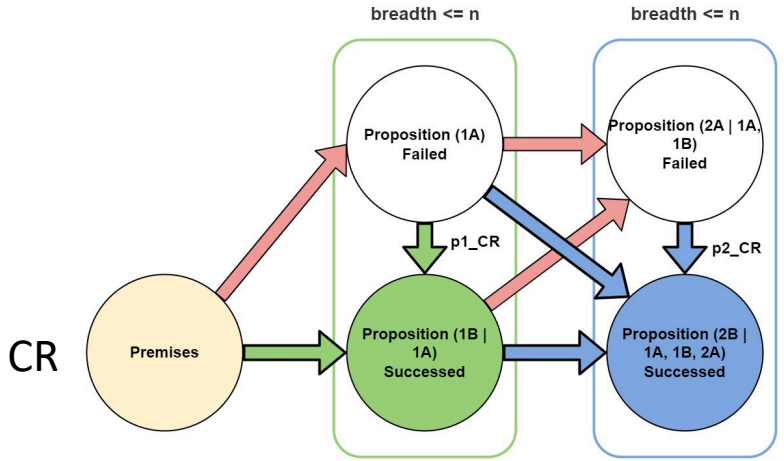
Why CR Works?





Suppose breadth = n,





$$p_{\mathrm{CoT}} \leq p_1 \cdot p_2$$

$$P_{\mathrm{CoT}} = 1 - (1 - p_{\mathrm{CoT}})^n \leq 1 - (1 - p_1 \cdot p_2)^n$$

$$P_{\mathrm{ToT}} = (1 - (1 - p_1)^n) \cdot (1 - (1 - p_2)^n)$$

$$P_{\mathrm{ToT}} \leq (1 - (1 - p_1 - q_1)^n) \cdot (1 - (1 - p_2 - q_2)^n) \leq P_{\mathrm{CR}}$$

$P_{CoT} \leq P_{ToT} \leq P_{CR}$

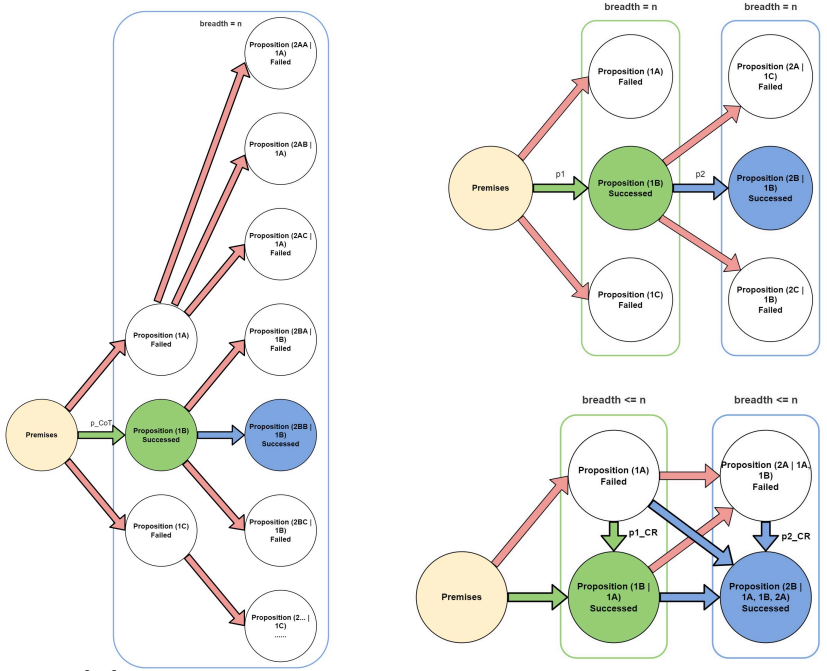
Suppose breadth = n,

$$p_{\mathrm{CoT}} \leq p_1 \cdot p_2$$

$$P_{\mathrm{CoT}} = 1 - (1 - p_{\mathrm{CoT}})^n \leq 1 - (1 - p_1 \cdot p_2)^n$$

$$P_{\mathrm{ToT}} = (1 - (1 - p_1)^n) \cdot (1 - (1 - p_2)^n)$$

$$P_{\mathrm{ToT}} \leq (1 - (1 - p_1 - q_1)^n) \cdot (1 - (1 - p_2 - q_2)^n) \leq P_{\mathrm{CR}}$$

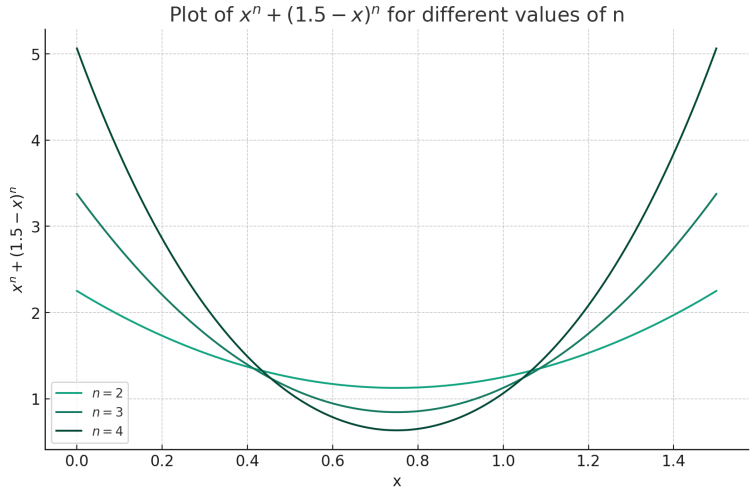


We only need to prove: $1 - (1 - p_1 \cdot p_2)^n \leq (1 - (1 - p_1)^n) \cdot (1 - (1 - p_2)^n)\,,\ \forall\, n \in \mathcal{Z}^{++}$ proposition

$$\Leftrightarrow\ 1 - (1 - p_1 \cdot p_2)^n \leq 1 - (1 - p_1)^n - (1 - p_2)^n + (1 - p_1)^n \cdot (1 - p_2)^n$$

$$\Leftrightarrow\ (1 - p_1)^n + (1 - p_2)^n \leq (1 - p_1 \cdot p_2)^n + (1 - p_1)^n \cdot (1 - p_2)^n$$

$$\Leftrightarrow\ (1 - p_1)^n + (1 - p_2)^n \leq (1 - p_1 \cdot p_2)^n + (1 - p_1 - p_2 + p_1 \cdot p_2)^n$$



$$(1 - p_1 \cdot p_2) + (1 - p_1 - p_2 + p_1 \cdot p_2) \equiv (1 - p_2) + (1 - p_2) \equiv 2 - p_1 - p_2$$

WLOG, let $p_1 \geq p_2$, then $(1 - p_1 - p_2 + p_1 \cdot p_2) \leq (1 - p_1) \leq (1 - p_2) \leq (1 - p_1 \cdot p_2)$

symmetry

From the right plot, it is easy to prove the proposition, hence we conclude that

$$P_{CoT} \leq P_{ToT} \leq P_{CR}$$

Thinking Fast and Slow

- **Dual modes of thinking**:
  - **System 1 (intuitive and fast) and System 2 (rational and slow).** **--- Daniel Kahneman**

- **Interaction between systems**: **System 1 (Proposer)** generates intuitions that are endorsed or corrected by **System 2 (Verifier)**, and pure **System 2 (Verifier)** can be lazy or depleted and needs intuitions from **System 1 (Proposer)**

- Verifiers can be implemented with symbolic systems or just LLMs

First-Order Logic Inference

Consider the following example adapted from the FOLIO dataset (Han et al., 2022), where empirically only the text statements (excluding logical propositions) will be given:

1. All monkeys are mammals: $\forall x(\text{Monkey}(x) \Rightarrow \text{Mammals}(x))$.
2. An animal is either a monkey or a bird: $\forall x(\text{Animal}(x) \Rightarrow (\text{Monkey}(x) \vee \text{Bird}(x)))$.
3. All birds fly: $\forall x(\text{Bird}(x) \Rightarrow \text{Fly}(x))$.
4. If something can fly, then it has wings: $\forall x(\text{Fly}(x) \Rightarrow \text{Wings}(x))$.
5. Rock is not a mammal, but Rock is an animal: $\neg\text{Mammal}(\text{Rock}) \wedge \text{Animal}(\text{Rock})$.

The question is: does Rock have wings? We have the following derivations:

a. The contrapositive of (1) is: $\forall x(\neg\text{Mammals}(x) \Rightarrow \neg\text{Monkey}(x))$.
b. (a) and (5) $\Rightarrow \neg\text{Monkey}(\text{Rock}) \wedge \text{Animal}(\text{Rock})$.
c. (2) and (5) $\Rightarrow (\text{Monkey}(\text{Rock}) \vee \text{Bird}(\text{Rock}))$
d. (b) and (c) $\Rightarrow \text{Bird}(\text{Rock})$.
e. (3) and (d) $\Rightarrow \text{Fly}(\text{Rock})$.
f. (4) and (e) $\Rightarrow \text{Wings}(\text{Rock})$.

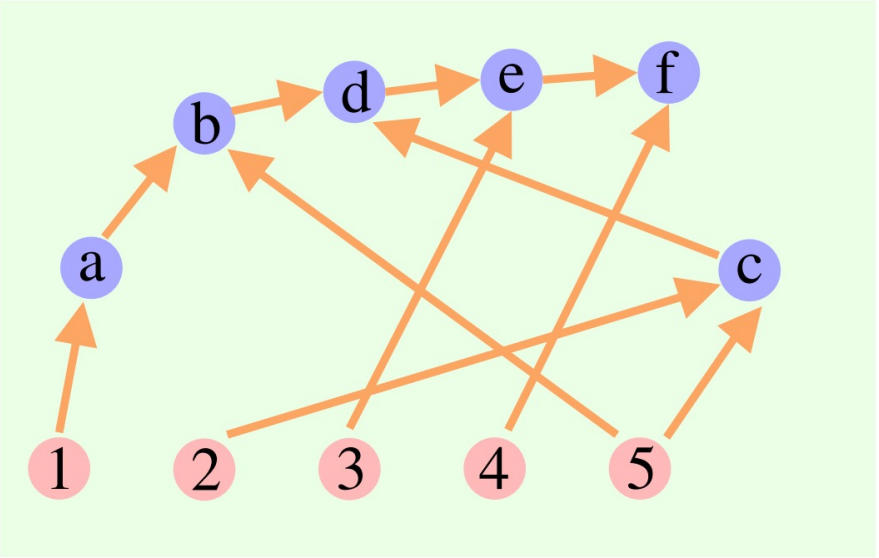


Figure 1: Illustration of our logical derivation

First-Order Logic Inference with CR

[Problem Description]

- Example ID: 687
- **Premises**:
  1. It is not true that some giant language models do not have good performance.
  2. All language models with good performance are used by some researchers.
  3. If a language model is used by some researchers, it is popular.
  4. If BERT is a giant language model, then GPT-3 is also a giant language model.
  5. BERT is a giant language model.
- **Hypothesis**: GPT-3 is popular.
- **Label**: **[True]**

[Chain-of-Thought Reasoning by GPT-4]

- **Reasoning**: "From the premises, we have:

  $$1.\neg(\exists x)(\text{Giant}(x) \land \neg\text{GoodPerformance}(x))$$
  $$2.\forall x(\text{GoodPerformance}(x) \rightarrow \text{UsedByResearchers}(x))$$
  $$3.\forall x(\text{UsedByResearchers}(x) \rightarrow \text{Popular}(x))$$
  $$4.\text{BERT} \rightarrow \text{GPT-3}$$
  $$5.\text{BERT}$$

  Now we know that BERT is a giant language model, so GPT-3 is also a giant language model (from premise 4 and 5). However, the premises do not provide information about GPT-3's performance or if it is used by researchers. Therefore, we cannot conclude if GPT-3 is popular."
- **Prediction**: **[Unknown]** (**Wrong**)

[Cumulative-Reasoning by GPT-4]

- **Generated Propositions**:
  1. "GPT-3 is a giant language model with good performance."
  2. "GPT-3 is used by some researchers."
- **Reasoning**: "We know that GPT-3 is a giant language model with good performance, and it is used by some researchers. According to the premises, if a language model is used by some researchers, it is popular. Therefore, GPT-3 is popular."
- **Prediction**: **[True]** (**Correct**)

Limitations of First-Order Logic Resolution

- **Premises:**
  - Alex was born on June 1st, 2000.
  - Alex is a British writer.
- **Hypothesis:** Alex was born on June 1st , 2001.

- Judge whether the Hypothesis is correct or not.

- **FOL Resolution gives [Unknown]**
- **It does not make sense!**

Limitations of FOL in Handling Natural Language Inference

[Problem Description]

- Example ID: 669
- **Premises**:
  1. Zaha Hadid is a British-Iraqi architect, artist and designer.
  2. Zaha Hadid was born on 31 October 1950 in Baghdad, Iraq.
  3. Hadid was a visiting professor of Architectural Design at the Yale School of Architecture.
  4. Max is an aspiring architecture student, and he plans to apply to Yale School of Architecture.
- **Hypothesis**: Hadid was born in 1982.
- **Label**: **[Unknown]** Wrong!
- **Explanation**: *We can see that Zaha Hadid was born on 31 October 1950 in Baghdad, Iraq. This directly contradicts the hypothesis that Hadid was born in 1982. It is common knowledge that people are born only once, and it is impossible for someone to be born in two different years.*

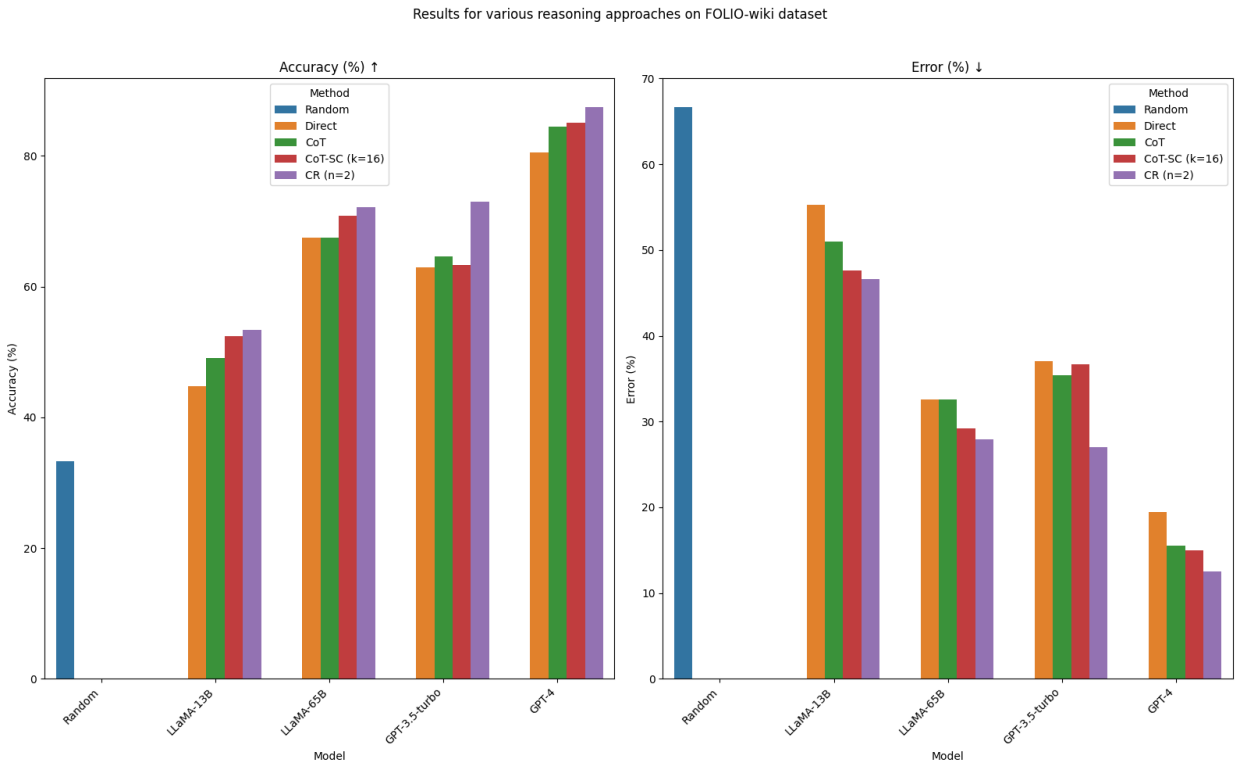
1. Missing common knowledge or contradictory to common knowledge; (9 in total, Example ID No. 34, 62, 162, 167, 228, 268, 526, 677, 679)
2. Overly ambiguous problems failing to provide unequivocal answers; (37 in total, Example ID No. 141, 215, 216, 223, 252, 261, 298, 321, 330, 396, 402, 409, 411, 431, 432, 456, 457, 482, 483, 496, 563, 572, 599, 624, 629, 641, 654, 660, 673, 682, 698, 750)
3. Inherent inconsistencies presented within the premises; (2 in total, Example ID No. 640, 643)
4. Vague premises or typographical errors; (2 in total, Example ID No. 314, 315)
5. Incorrect answers. (24 in total, Example ID No. 9, 46, 52, 84, 100, 144, 273, 276, 299, 310, 322, 345, 367, 437, 452, 453, 464, 557, 573, 578, 605, 632, 671, 715)

Around 13.8% of the FOLIO wiki dataset

Results on FOLIO wiki dataset

Table 1: Results for various reasoning approaches on FOLIO-wiki dataset.

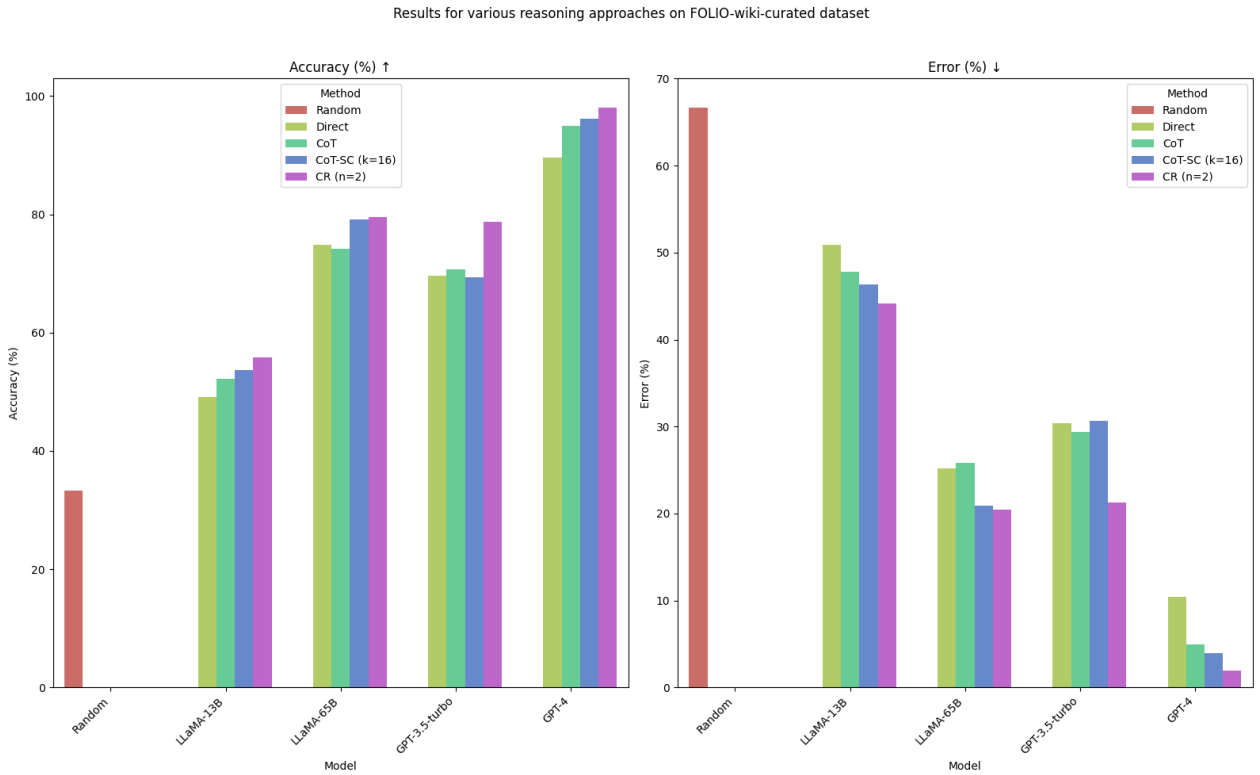
Model	Method	Acc. ↑ (%)	Error ↓ (%)
-	[Random]	33.33	66.67
LLaMA-13B	Direct	44.75	55.25
	CoT	49.06 (+4.31)	50.94 (-4.31)
	CoT-SC (k = 16)	52.43 (+7.68)	47.57 (-7.68)
	CR (**ours**, n = 2)	**53.37** (+8.62)	**46.63** (-8.62)
LLaMA-65B	Direct	67.42	32.58
	CoT	67.42 (+0.00)	32.58 (-0.00)
	CoT-SC (k = 16)	70.79 (+3.37)	29.21 (-3.37)
	CR (**ours**, n = 2)	**72.10** (**+4.68**)	**27.90** (**-4.68**)
GPT-3.5-turbo	Direct	62.92	37.08
	CoT	64.61 (+1.69)	35.39 (-1.69)
	CoT-SC (k = 16)	63.33 (+0.41)	36.67 (-0.41)
	CR (**ours**, n = 2)	**73.03** (**+10.11**)	**26.97** (**-10.11**)
GPT-4	Direct	80.52	19.48
	CoT	84.46 (+3.94)	15.54 (-3.94)
	CoT-SC (k = 16)	85.02 (+4.50)	14.98 (-4.50)
	CR (**ours**, n = 2)	**87.45** (**+6.93**)	**12.55** (**-6.93**)



Results on FOLIO wiki curated dataset

Table 2: Results for various reasoning approaches on FOLIO-wiki-curated dataset.

Model	Method	Acc. ↑ (%)	Error ↓ (%)
-	[Random]	33.33	66.67
LLaMA-13B	Direct	49.13	50.87
	CoT	52.17 (+3.04)	47.83 (-3.04)
	CoT-SC (k = 16)	53.70 (+4.57)	46.30 (-4.57)
	CR (**ours**, n = 2)	**55.87** (**+6.74**)	**44.13** (**-6.74**)
LLaMA-65B	Direct	74.78	25.22
	CoT	74.13 (-0.65)	25.87 (-0.65)
	CoT-SC (k = 16)	79.13 (+4.35)	20.87 (-4.35)
	CR (**ours**, n = 2)	**79.57** (**+4.79**)	**20.43** (**-4.79**)
GPT-3.5-turbo	Direct	69.57	30.43
	CoT	70.65 (+1.08)	29.35 (-1.08)
	CoT-SC (k = 16)	69.32 (-0.25)	30.68 (+0.25)
	CR (**ours**, n = 2)	**78.70** (**+9.13**)	**21.30** (**-9.13**)
GPT-4	Direct	89.57	10.43
	CoT	95.00 (+5.43)	5.00 (-5.43)
	CoT-SC (k = 16)	96.09 (+6.52)	3.91 (-6.52)
	CR (**ours**, n = 2)	**98.04** (**+8.47**)	**1.96** (**-8.47**)



Limitations of First-Order Logic Resolution

- **Limitations of Expressiveness:** First-order logic (FOL) lacks the expressive power to capture some properties of the real numbers. For example, properties involving uncountably many real numbers often cannot be expressed in FOL. In addition, properties requiring quantification over sets of real numbers or functions from real numbers to real numbers cannot be naturally represented in FOL.

- **Translation Misalignment**: Risk of semantic discrepancies during translation, rendering resolutions ineffective. For instance, translating statements as $\forall x\ Bird(x) \Rightarrow CanFly(x)$ and $\forall x\ \big(Fly(x) \Rightarrow Wings\ (x)\big)$ may cause a misalignment between "CanFly" and "Fly", leading to flawed conclusions. It often fails to capture the full richness and ambiguity of natural language. *(see https://en.wikipedia.org/wiki/First-order_logic#Limitations)*

- **Undecidability**: The general problem of determining the truth of a statement in FOL is undecidable, constraining its applicability for automated reasoning in complex systems.
  (connected to halting problem: http://kilby.stanford.edu/~rvg/154/handouts/fol.html)

Higher-Order Logic Inference with CR

[Modified Example FraCas-317]

- **Premises**:
  1. Most of the representatives who read the report have a positive attitude towards it.
  2. No two representatives have read it at the same time, and they may have different opinions about it.
  3. No representative took less than half a day to read the report.
  4. There are sixteen representatives.
- **Hypothesis**: It took the representatives more than a week to read the report, and most found it valuable.
- **Label**: **[True]**
- **Higher-Order Logic Premises**:
  1. $\text{most}(\lambda x.\text{representative}(x) \wedge \text{reads}(x, \text{report}), \lambda x.\text{has\_positive\_attitude}(x, \text{report}))$
  2. $$\neg\exists x, y\,(x \neq y \wedge \text{representative}(x) \wedge \text{representative}(y) \wedge \text{read\_at\_same\_time}(x, y, \text{report})) \wedge \forall x, y\,(x \neq y \wedge \text{representative}(x) \wedge \text{representative}(y) \rightarrow \text{may}(\lambda z.\text{different\_opinions}(z, x, y, \text{report})))$$
  3. $\neg\exists x(\text{representative}(x) \wedge \text{took\_less\_than\_half\_day}(\lambda y.\text{read}(y, \text{report})))$
  4. $\text{count}(\lambda x.\text{representative}(x)) = 16$
- **Higher-Order Logic Hypothesis**:

$$\exists t(\text{total\_reading\_time}(\text{report}) > 7 \wedge \text{time\_unit}(t, \text{day})) \wedge \text{most}(\lambda x.\text{representative}(x), \lambda x.\text{found\_valuable}(x, \text{report}))$$

[Cumulative-Reasoning by GPT-4]

- **Generated Propositions**:
  1. "Using P1 and P2, we can deduce Q1: Every representative read the report at a different time, and most representatives found the report valuable."
  2. "Using P3 and Q1, we deduce Q2: The total reading time must be at least 8 days. (since 16 representatives each took at least half a day)."
  3. "Using Q2 and P4, we deduce Q3: The representatives took at least 8 days to read the report."
- **Reasoning**: "Based on the intermediate propositions Q1, Q2, and Q3, the hypothesis that it took the representatives more than a week to read the report, and most found it valuable, is confirmed as True."
- **Prediction**: **[True]** **(Correct)**

Crafting higher-order logic programs that are solvable by symbolic systems is a daunting task, even for experts.

It is also challenging for Large Language Models (LLMs) to write these intricate programs effectively.

Using Formal Theorem Provers based on Higher-order (Categorical) Logic and (Dependent) Type Theory ups the ante, making it exponentially harder.

However, CR solves these problems pretty well without resorting to symbolic systems, just like the way humans think.

AutoTNLI dataset

Janet Leigh (Original)		
Born	July 6, 1927	
Died	October 3, 2004	
Children	Kelly Curtis; Jamie Lee Curtis	
Alma Mater	Stanford University	
Occupation	None	
H1: Janet Leigh was born before 1940.		E
H2: The age of Janet Leigh is more than 70.		E
H3: Janet Leigh has 1 children		C
H4: Janet Leigh graduated from Stanford University		E

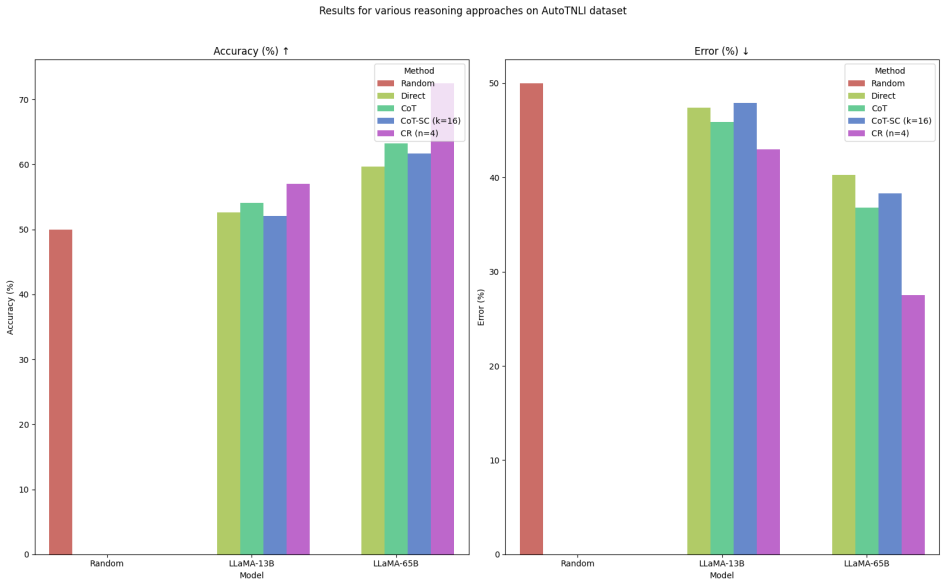
- Temporal Logic
- Mathematical Reasoning
- Common Knowledge

Results on AutoTNLI dataset

Evaluation Results. As shown in Table 3, both LLaMA-13B and LLaMA-65B models reveal that CR delivers a significant enhancement in performance compared to CoT, with a relative improvement reaching up to 9.3% on the LLaMA-65B model. This data emphasizes the clear advantage of CR over CoT and CoT-SC techniques in the framework of the AutoTNLI dataset.

Table 3: Results for various reasoning approaches on AutoTNLI dataset.

Model	Method	Acc. ↑ (%)	Error ↓ (%)
-	[Random]	50.00	50.00
LLaMA-13B	Direct	52.6	47.4
	CoT	54.1 (+1.5)	45.9 (-1.5)
	CoT-SC (k = 16)	52.1 (-0.5)	47.9 (+0.5)
	CR (**ours**, $n = 4$)	**57.0** (**+5.4**)	43.0 (**-5.4**)
LLaMA-65B	Direct	59.7	40.3
	CoT	63.2 (+3.5)	36.8 (-3.5)
	CoT-SC (k = 16)	61.7 (+2.0)	38.3 (-2.0)
	CR (**ours**, $n = 4$)	**72.5** (**+12.8**)	27.5 (**-12.8**)



Game of 24

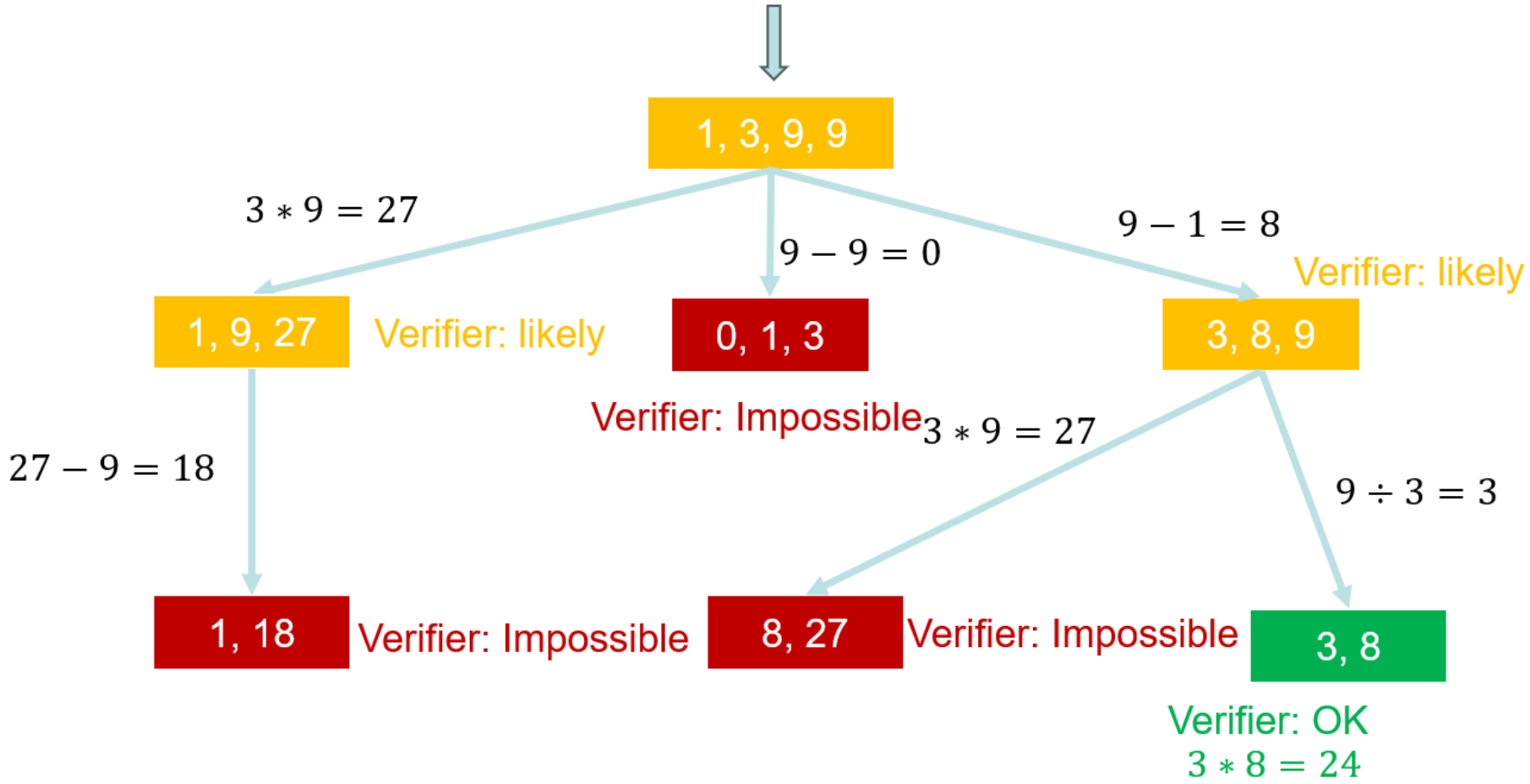
- Combine four specified integers using basic arithmetic operations to get the number 24.

[Illustrative example for Game of 24]

- Numbers: [3, 3, 7, 7]
- Arithmetic Operations: [+, −, ×, /, (,)]
- **Solution**:

$$(3 + 3/7) \times 7 = 24$$

CR on Game of 24

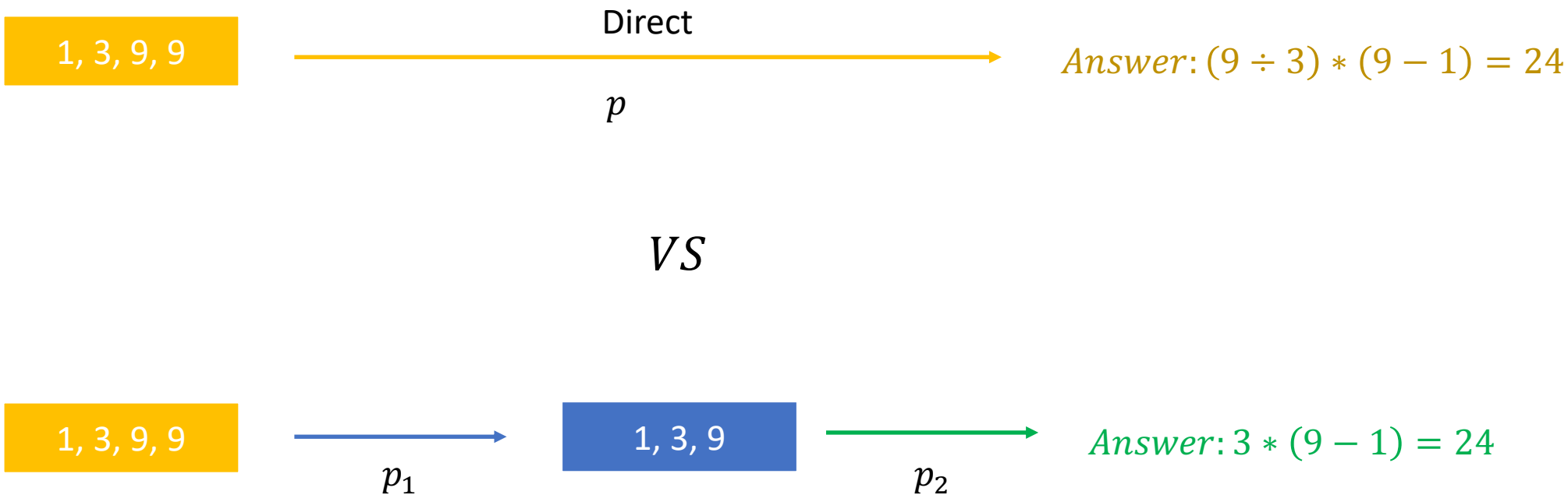


Compare with ToT

- Each iteration, generates at most one newly reached state.

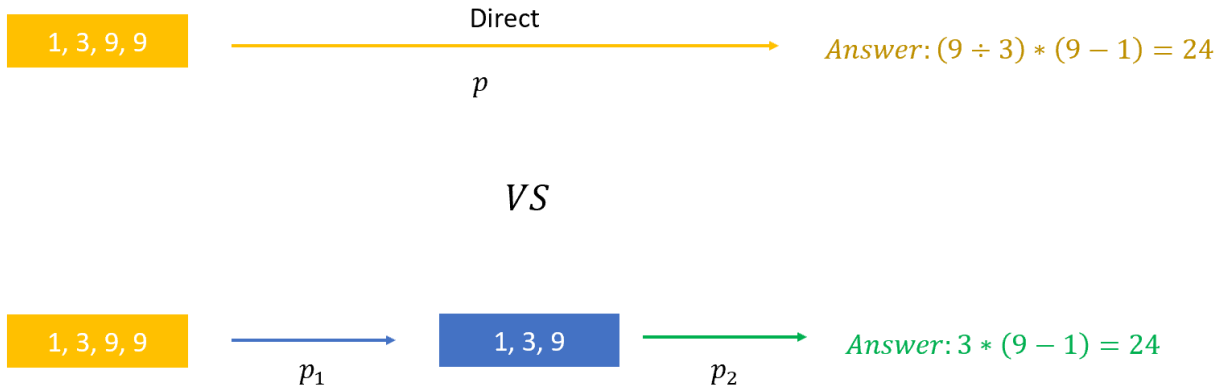
- CR allows the LLM to determine the search depth autonomously, and performs different search widths on different layers of the search tree.

Why CR?


1, 3, 9, 9
Direct
p
$Answer: (9 \div 3) * (9 - 1) = 24$
VS
1, 3, 9, 9
p_1
1, 3, 9
p_2
$Answer: 3 * (9 - 1) = 24$

Why CR?

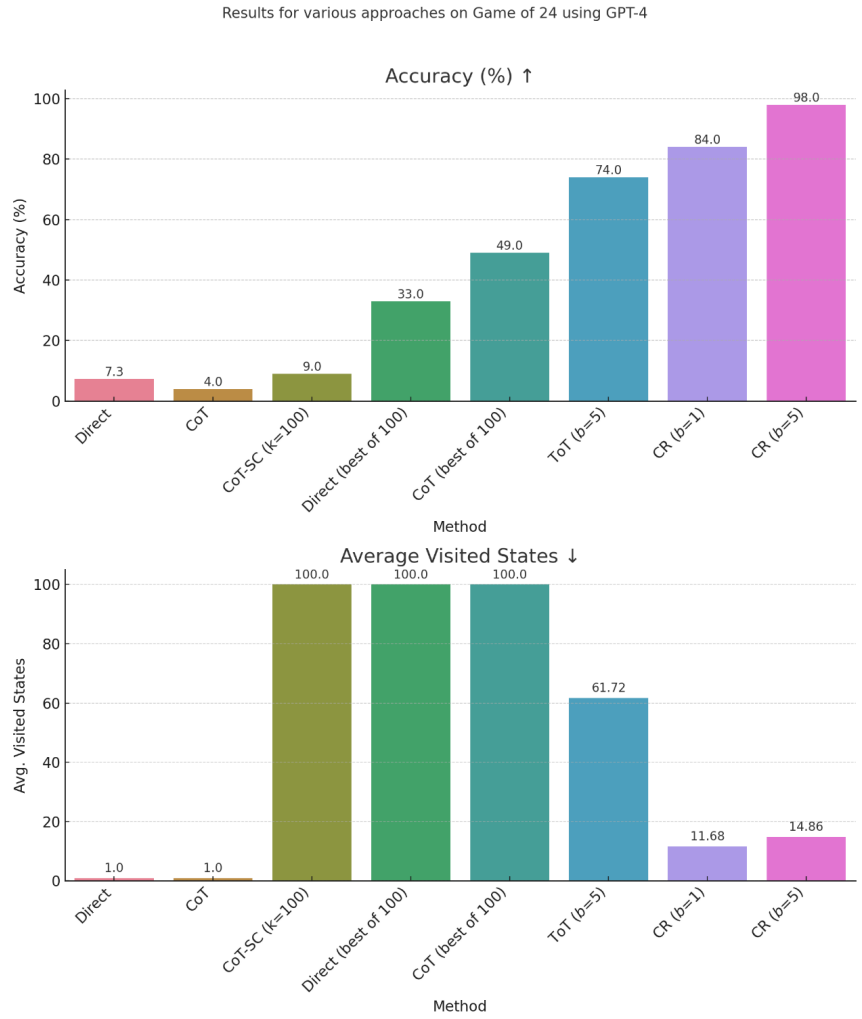
Puzzle	p (%)	p_1 (%)	p_2 (%)	p_1p_2 (%)
2,7,12,13	3.0	62.3	8.0	5.0 (+2.0)
6,11,12,13	0.0	64.8	8.0	5.2 (+5.2)
8,8,10,12	1.8	6.9	63.9	4.4 (+2.6)



Results on Game of 24

Table 4: Results for various approaches on Game of 24 using GPT-4. The average number of visited states for ToT is computed from the experimental logs available in its official GitHub repository.

Method	Acc. ↑ (%)	# Avg. visited states ↓
Direct	7.3	1
CoT	4.0	1
CoT-SC (k = 100)	9.0	100
Direct (best of 100)	33	100
CoT (best of 100)	49	100
ToT ($b = 5$)	74	61.72
CR (ours, $b = 1$)	84 (+10)	**11.68 (-50.04)**
CR (ours, $b = 2$)	94 (+20)	13.70 (-48.02)
CR (ours, $b = 3$)	97 (+23)	14.25 (-47.47)
CR (ours, $b = 4$)	97 (+23)	14.77 (-46.95)
CR (ours, $b = 5$)	**98 (+24)**	14.86 (-46.86)



Mathematical Reasoning on MATH with CR

[Problem Description]

- Example ID: test/intermediate_algebra/1350.json
- Level: 5
- Subject: Intermediate Algebra
- **Problem**: Consider the polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

where the polynomial has integer coefficients and its roots are distinct integers.
Given $a_n = 2$ and $a_0 = 66$, the inquiry is to determine the least possible value of $|a_{n-1}|$.

[Ground Truth Solution]

- **Solution**: Since $f(x)$ has integer coefficients, the Integer Root Theorem asserts that any integer roots of $f(x)$ must divide the constant term $66 = 2 \cdot 3 \cdot 11$. Consequently, the potential integer roots of $f(x)$ are

$$\pm 1, \ \pm 2, \ \pm 3, \ \pm 6, \ \pm 11, \ \pm 22, \ \pm 33, \ \pm 66.$$

Additionally, given that all roots of $f(x)$ are integers, they are necessarily members of the aforementioned list.
We proceed to utilize Vieta's formulas. The roots of $f(x)$ yield a product of $(-1)^n \cdot \frac{a_0}{a_n}$, which evaluates to either 33 or -33. Simultaneously, the sum of these roots is $-\frac{a_{n-1}}{a_n} = -\frac{a_{n-1}}{2}$. To minimize $|a_{n-1}|$, we aim to reduce the absolute value of the root sum, ensuring that the product of the roots remains 33 or -33.
Considering two distinct scenarios:
Case 1: One of the roots is 33 or -33. In this scenario, the only other viable roots are ± 1. Here, the root sum's absolute value is at least 32.
Case 2: One root is 11 or -11, and another is 3 or -3. The only other plausible roots remain ± 1, leading to a root sum's absolute value of at least $11 - 3 - 1 = 7$. This is a more optimal outcome than Case 1. For an absolute root sum of 7, we find $|a_{n-1}| = 7|a_n| = 14$. From our analysis, we deduce that $|a_{n-1}| \geq 14$. Verifying this, we observe that equality is attained by

$$\begin{aligned} f(x) &= 2(x+11)(x-3)(x-1) \\ &= 2x^3 + 14x^2 - 82x + 66, \end{aligned}$$

which is consistent with having integer coefficients and integer roots. Thus, the minimal achievable value for $|a_{n-1}|$ is $\boxed{14}$.

- **Answer**: **14**

Figure 4: An example from the MATH dataset (Hendrycks et al., 2021; Lightman et al., 2023).

[Chain-of-Thought Reasoning by GPT-4]

- **Reasoning**: "Let's think step by step:
  1. The roots of the polynomial are distinct integers, and the leading coefficient is 2, so the roots must be divisors of $\frac{66}{2} = 33$.
  2. The divisors of 33 are $\pm 1, \pm 3, \pm 11, \pm 33$.
  3. The sum of the roots of the polynomial is $-\frac{a_{n-1}}{a_n}$.
  4. To minimize $|a_{n-1}|$, we should select the roots to be as close to zero as possible.
  5. The smallest possible sum of the roots is $-1 - 3 + 11 + 33 = 40$,
  6. Thus, the smallest possible value of $|a_{n-1}|$ is $2 \times 40 = \boxed{80}$."
- **Answer**: **80** (Wrong)

[Cumulative-Reasoning by GPT-4]

- **Generated Hints**:
  1. "**Hint 1**: Recall Vieta's formulas, which relate the coefficients of a polynomial to the sums and products of its roots."
  2. "**Hint 2**: The product of the roots of the polynomial is equal to $-\frac{a_0}{a_n}$."
  3. "**Hint 3**: The sum of the roots of the polynomial is equal to $-\frac{a_{n-1}}{a_n}$."
  4. "**Hint 4**: Since the roots are distinct integers, consider the factors of $-\frac{a_0}{a_n}$ and their sums."
- **Generated Simple Questions and Answers**:

Question 1: "What is the product of the roots of the polynomial $f(x)$?"
Answer 1: "The product of the roots of the polynomial is $-\frac{a_0}{a_n} = -\frac{66}{2} = -33$."
Question 2: "What are the possible sets of distinct integer roots of the polynomial?"
Answer 2: "The possible sets of distinct integer roots are $\{-33, 1\}$, $\{-11, 3\}$, and $\{-3, 11, -1\}$."
Question 3: "What are the sums of the roots for each possible set of roots?"
Answer 3: "The sums of the roots are -32, -8, and 7 for the sets $\{-33, 1\}$, $\{-11, 3\}$, and $\{-3, 11, -1\}$, respectively."
Question 4: "What is the least possible value of $|a_{n-1}|$?"
Answer 4: "The least possible value of $|a_{n-1}|$ is $2 \times 7 = 14$."

- **Solution**: "Let's think step by step. Given that $a_n = 2$ and $a_0 = 66$, and the roots of the polynomial are distinct integers, the least possible value of $|a_{n-1}|$ is $\boxed{14}$."
- **Answer**: **14** (Correct)

Figure 5: Solutions for the example Fig. 4 from the MATH dataset, generated by CoT and CR. CoT will generate the answer directly through a chain of thought. By contrast, CR will first generate a few hints, then several simple and foundational questions, and then answer them by self, and finally conclude with the help of the generated hints and question-answer pairs.

Mathematical Reasoning on MATH with CR

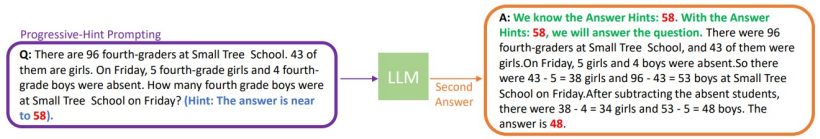
Table 5: Comparative performance on the MATH dataset using GPT-4. We adopted a default temperature setting of $t = 0.0$, consistent with prior research settings (greedy decoding). PHP denotes the application of the progressive-hint prompting. "Iters" represents the average number of LLM interactions, and **Overall** reflects the overall results across MATH subtopics.

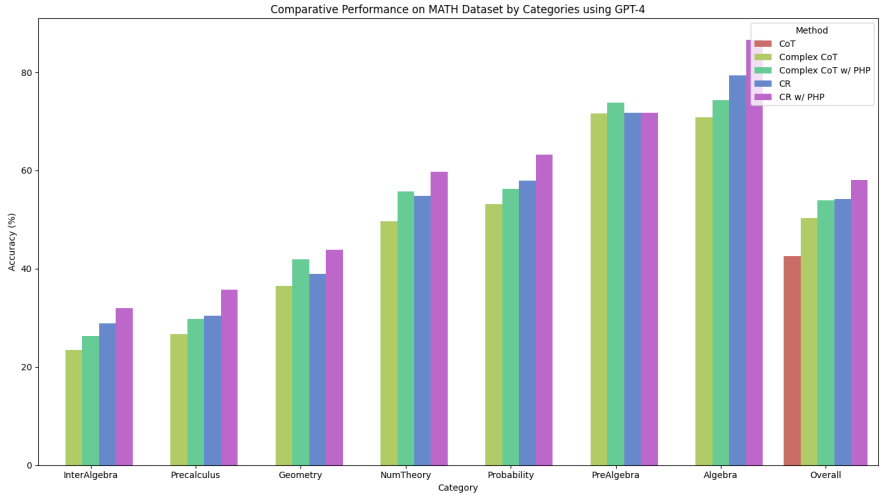
	w/ PHP	MATH Dataset (* denotes using 500 test examples subset following Lightman et al. (2023))							
		InterAlgebra	Precalculus	Geometry	NumTheory	Probability	PreAlgebra	Algebra	**Overall**
CoT (OpenAI, 2023)	✗	-	-	-	-	-	-	-	42.50
Complex CoT, 8-shot (Zheng et al., 2023)	✗	23.4	26.7	36.5	49.6	53.1	71.6	70.8	50.36
	✓	26.3	29.8	41.9	55.7	56.3	73.8	74.3	53.90
	(Iters)	3.2414	3.2435	3.2233	3.1740	2.8122	2.3226	2.4726	2.8494
Complex CoT* (repro., 8-shot)	✗	29.9	33.9	34.1	46.8	47.4	62.1	70.7	48.80
	✓	28.9	30.4	43.9	53.2	50.0	68.5	84.1	53.80
	(Iters)	2.7629	2.4643	2.7805	2.7581	2.4474	2.3780	2.5484	2.59
CR* (**ours**, 4-shot)	✗	28.9 (-1.0)	30.4 (-3.5)	39.0 (+4.9)	54.8 (+8.0)	57.9 (+10.5)	71.8 (+9.7)	79.3 (+8.6)	**54.20** (+5.40)
	✓	**32.0** (+3.1)	**35.7** (+5.3)	**43.9** (+0.0)	**59.7** (+6.5)	**63.2** (+13.2)	**71.8** (+3.3)	**86.6** (+2.5)	**58.00** (+4.20)
	(Iters)	2.6598	2.4821	2.5122	2.2903	2.2105	2.2195	2.3548	**2.40** (-0.19)

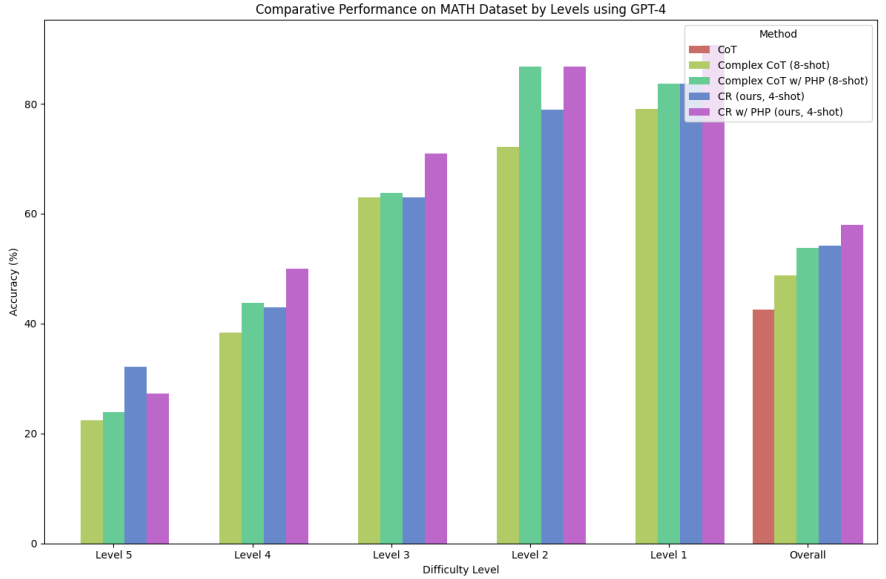
Table 6: Comparative performance on the MATH dataset using GPT-4 for different difficulty levels.

	w/ PHP	MATH Dataset (* denotes using 500 test examples subset)					
		Level 5	Level 4	Level 3	Level 2	Level 1	**Overall**
CoT (OpenAI, 2023)	✗	-	-	-	-	-	42.50
Complex CoT* (repro., 8-shot)	✗	22.4	38.3	62.9	72.2	79.1	48.80
	✓	23.9	43.8	63.8	86.7	83.7	53.80
CR* (**ours**, 4-shot)	✗	**32.1** (+9.7)	43.0 (+4.7)	62.9 (+0.0)	78.9 (+6.7)	83.7 (+4.6)	**54.20** (+5.40)
	✓	27.3 (+3.4)	**50.0** (+6.2)	**70.9** (+7.1)	**86.7** (+0.0)	**90.7** (+7.0)	**58.00** (+4.20)

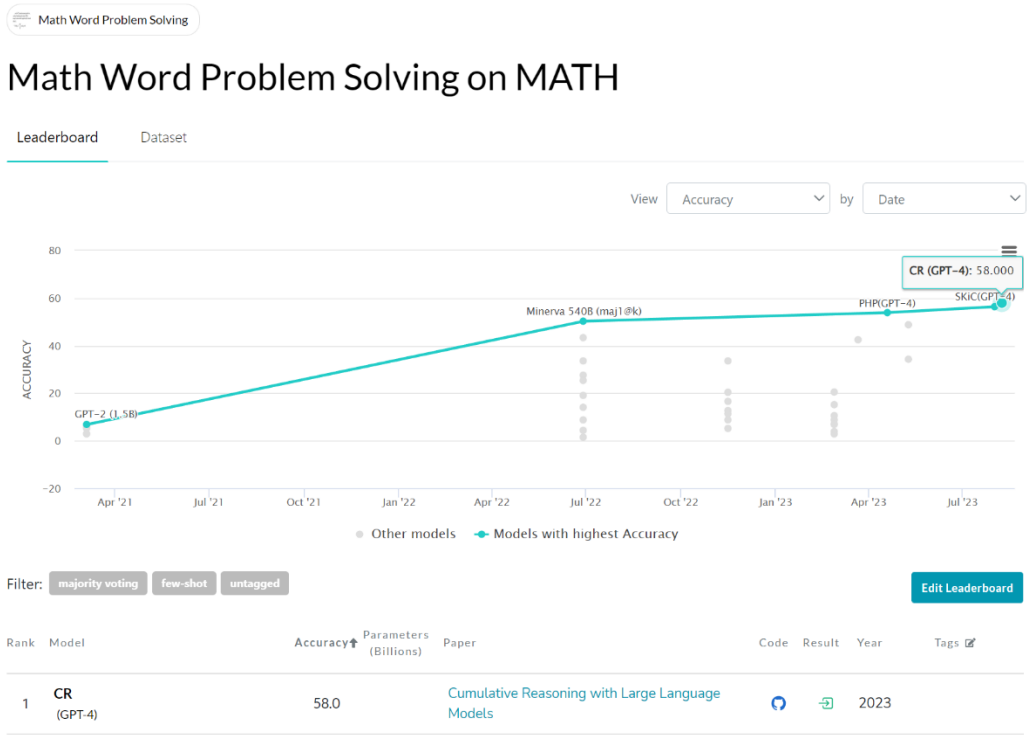
+42%!







Mathematical Reasoning on MATH with CR



For official implementations, please refer to:
https://github.com/iiis-ai/cumulative-reasoning

Thanks!