arXiv:2308.04371v4 [cs.AI] 25 Aug 2023

Cumulative Reasoning With Large Language Models

Yifan Zhang[*1] **Jingqin Yang**[*1] **Yang Yuan**[1,2] **Andrew Chi-Chih Yao**[1,2]

[1]IIIS, Tsinghua University
[2]Shanghai Qizhi Institute
{yuanyang,andrewcyao}@tsinghua.edu.cn

Abstract

While language models are powerful and versatile, they often fail to address highly complex problems. This is because solving complex problems requires deliberate thinking, which has been only minimally guided during training. In this paper, we propose a new method called Cumulative Reasoning (CR), which employs language models in a cumulative and iterative manner to emulate human thought processes. By decomposing tasks into smaller components, CR streamlines the problem-solving process, rendering it both more manageable and effective. For logical inference tasks, CR consistently outperforms existing methods with an improvement up to 9.3%, and achieves the astonishing accuracy of 98.04% on the curated FOLIO wiki dataset. In the context of the Game of 24, CR achieves an accuracy of 98%, which signifies a substantial enhancement of 24% over the previous state-of-the-art method. Finally, on the MATH dataset, we establish new state-of-the-art results with 58.0% overall accuracy, surpassing the previous best approach by a margin of 4.2%, and achieving 43% relative improvement on the hardest level 5 problems (22.4% → 32.1%) [†].

1 Introduction

Despite the remarkable advances made by large language models (LLMs) in a variety of applications (Devlin et al., 2018; Radford et al., 2018, 2019; Brown et al., 2020; Raffel et al., 2020; OpenAI, 2023), they still struggle to provide stable and accurate answers when faced with highly complex tasks. For instance, it has been observed that language models have difficulty directly generating correct answers for high school math problems (Lightman et al., 2023).

This shortfall may be anticipated, considering the training approach adopted by LLMs. Specifically, they are trained to sequentially predict the next token based on the given context, without a pause for deliberate thoughts. As elucidated by Kahneman (2011), our cognitive processing processes comprise two distinct systems: System 1 is fast, instinctive, and emotional; System 2 is slow, deliberate, and logical. 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. These include the Chain-of-Thought (CoT) that prompts the model to offer step-by-step solutions (Wei et al., 2022), and the Tree-of-Thought (ToT) that 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.

---

[*] Equal Contribution.

[†] The code is available at `https://github.com/iiis-ai/cumulative-reasoning`.

Preprint. Under review.

In this paper, we propose a new method termed Cumulative Reasoning (CR), which presents a more general characterization of the thinking process. CR employs three distinct LLMs: the proposer, verifier, and reporter. The proposer keeps proposing potential propositions, which were verified by one or more verifiers, and the reporter decides when to stop and report the solution.

CR significantly amplifies the power of language models in addressing complex tasks, achieved by decomposing each task into atomic and manageable steps. Despite the computational infeasibility of enumerating the exponentially numerous possible complex tasks, CR ensures that each individual step can be efficiently learned and resolved. This strategic decomposition effectively transforms an otherwise unmanageable exponential problem into a sequence of solvable tasks, thereby providing a robust solution to the original problem.

Our empirical analyses include three components. In the first experiment, we tackled logical inference tasks like FOLIO wiki (pertaining to first-order logic) and AutoTNLI (associated with higher-order logic). On these datasets, CR consistently surpassed current methodologies, showcasing an enhancement of up to 9.3%. Additionally, a rigorous refinement of the FOLIO dataset generated the "FOLIO wiki curated," on which CR recorded a remarkable accuracy of 98.04%. In the second experiment, which revolved around the Game of 24, CR achieved an accuracy of 98%. Remarkably, this represents a significant improvement of 24% when compared to the prior state-of-the-art method, ToT (Yao et al., 2023). In the last experiment, we established new state-of-the-art results on the renowned MATH dataset (Hendrycks et al., 2021), achieving 58.0% overall accuracy with a margin of 4.2% over the Complex-CoT with PHP method (Fu et al., 2022; Zheng et al., 2023). Noteworthy, our method achieves 43% relative improvement on the hardest level 5 problems (22.4% $\to$ 32.1%).

2 Preliminaries

2.1 Propositional logic

Propositional logic, the most fundamental system of logic, encompasses elements p, q, r and a variety of operations. These include "and" ($p \wedge q$), "or" ($p \vee q$), "implies" ($p \Rightarrow q$), and "not" ($\neg p$). The constants true and false are denoted as 1 and 0 respectively. This system adheres to the following rules:

$$x \wedge x = x, \quad x \vee x = x, \quad 1 \wedge x = 1, \quad 0 \vee x = 0, \quad x \wedge (y \vee x) = x = (x \wedge y) \vee x.$$

and distributive laws:

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z), \quad x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z).$$

In a Boolean algebra, every element x has a complement $\neg x$ and the following holds true:

$$x \wedge \neg x = 0, \quad x \vee \neg x = 1, \quad \neg\neg x = x.$$

2.2 Higher-order logic

Building upon propositional logic, first-order logic (FOL) introduces universal quantification ($\forall$) and existential quantification ($\exists$) to describe more intricate propositions. For instance, the statement "$\forall_x \mathrm{Dog}(x) \Rightarrow \mathrm{Animal}(x)$" translates to "for every x, if x is a dog, then it is also an animal".

Higher-order logic (HOL) represents a sophisticated formalism that permits quantification over functions and predicates, an ability that contrasts sharply with FOL, which restricts quantification to individual objects. The distinctive characteristics of HOL, as opposed to FOL, can be elaborated as follows (Mineshima et al., 2015):

Quantification over Functions: Higher-order logic (HOL) allows for lambda expressions, such as $\lambda y.\mathrm{report\_attribute}(y, \mathrm{report})$, whereby functions themselves become the subject of quantification. An illustration of this is found in the expression "a representative who reads this report." Here, quantification spans the predicates representing both the representative and the reading of the report, a phenomenon captured as a higher-order function. Unlike HOL, FOL is incapable of extending quantification to functions or predicates.

Generalized Quantifiers: The introduction of generalized quantifiers, such as "most," serves as another demarcation line between HOL and FOL. These quantifiers are capable of accepting predicates as arguments, enabling the representation of relations between sets, a feat that transcends the expressive capacity of FOL.

Modal Operators: Employing modal operators like "might" signifies a transition towards HOL. These operators, applicable to propositions, give rise to multifaceted expressions that defy easy reduction to the confines of FOL.

Attitude Verbs and Veridical Predicates: The integration of attitude verbs, such as "believe," and veridical predicates like "manage," injects an additional layer of complexity necessitating the use of HOL. These linguistic constructs can engage with propositions as arguments, interacting with the truth values of those propositions in subtle ways that demand reasoning extending beyond the capabilities of FOL.

2.3 Illustrative example

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:

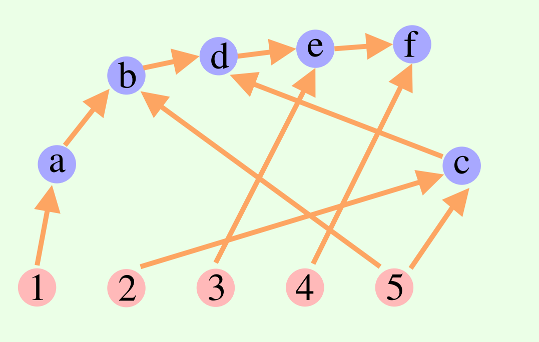

Figure 1: Illustration of our logical derivation

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) $\Rightarrow \text{Wings}(\text{Rock})$.

While the derivation can be treated as a general "chain of thought" from (a) to (f), its internal structure is neither a chain nor a tree. Instead, it is a directed acyclic graph (DAG), with each directed edge as one step of derivation. For examples of higher-order logic, see Appendix A.

3 Our Method

3.1 Cumulative Reasoning (CR)

Our CR algorithm 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.

See Figure 2 for an illustration. In each iteration, the proposer initiates the process by proposing one or a few new claim(s) based on existing predicates. Subsequently, the verifier(s) evaluate the proposal, determining whether the claim(s) can be retained as a new predicate. Finally, the reporter decides if it is the optimal time to cease the thought process and deliver the answer.

Ideally, the proposer should be implemented using a language model pretrained on the corresponding derivation tasks. Verifier(s) should be capable of translating the derivations to appropriate formal systems and verifying them using symbolic reasoning modules such as a propositional logic solver or a formal math prover. However, one can also use general foundational models like GPT-4 or LLaMA, with different prompts for these roles.

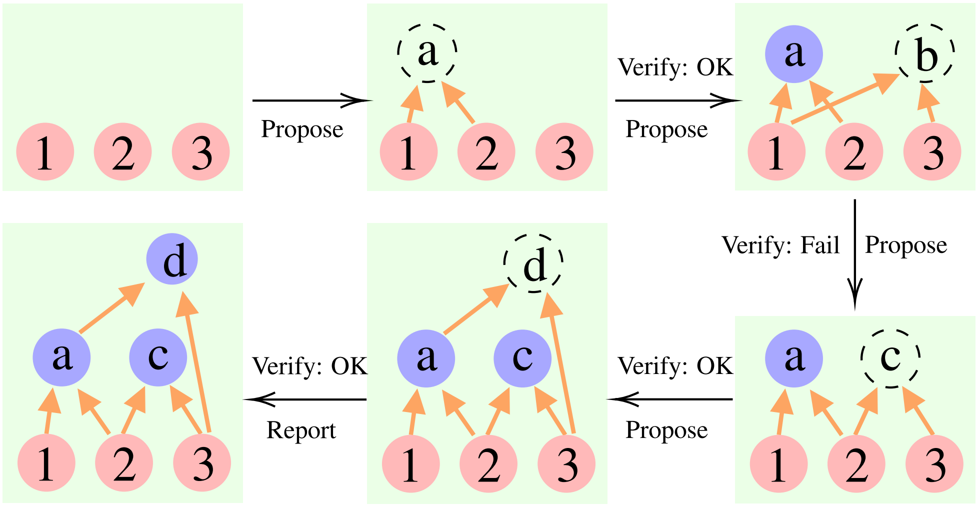

Figure 2: An illustration of CR Reasoning for a 3-premises problem.

3.2 Compare with CoT and ToT

CR clearly generalizes CoT (Wei et al., 2022), in the sense that if there are no verifiers, and the proposer keeps proposing the next steps until the end, CR becomes the standard chain of thought. However, in CR the overall thinking process is not necessarily a chain or a tree, it can be a DAG. Therefore, CR can be used for solving more complex problems.

At first glance, CR is similar to the ToT, which solves the problems with a thought search tree (Yao et al., 2023; Long, 2023). However, our method is more general in the sense that it stores all the historical correct reasoning results in memory, which can be a DAG. By contrast, ToT will not store the information from other branches for exploration at the current search branch.

4 Experiments

Our experimental framework is based on the Microsoft guidance library (Lundberg et al., 2023), which offers the flexibility to intertwine generation, prompting, and logical control in a seamless flow that aligns with language models. We consider the following LLMs:

- GPT-3.5-turbo
- GPT-4
- LLaMA-13B

- LLaMA-65B

Our Proposer, Verifier(s), and Reporter in CR are implemented using the same LLM with different few-shot prompts. This approach ensures a broad application scope and simplifies implementation. For optimal results, future work could consider the application of a Proposer pretrained on task-specific corpus and Verifier(s) aided by symbolic formal systems. We denote n as the number of generated intermediate propositions, and k as the number of majority voting times. We set the temperature $t = 0.1$ by default and $t = 0.7$ for majority voting. We also remark that both gpt-3.5-turbo and gpt-4 operate as chat-format APIs from OpenAI.

[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:

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

  Now we know that BERT is a giant language model, so GPT-3 is also a giant language model (from premises 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**)

Figure 3: An example from the FOLIO dataset, with solutions generated by CoT and CR. CoT will generate the answer directly through a chain of thought. By contrast, CR will first generate a few propositions, and conclude with the help of the generated propositions.

4.1 FOLIO wiki

FOLIO dataset (Han et al., 2022) is a first-order logical inference dataset for reasoning in natural language. The label of each problem can be "True", "False", or "Unknown". See Figure 3 for an example. We observed that while the Chain-of-Thought reasoning process can generate useful intermediary results, it tends to flounder midway, failing to arrive at the correct conclusion. Conversely, the CR initially spawns two beneficial propositions and leverages them to successfully solve the

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)**

problem at hand. For a deeper dive into specific examples of the FOLIO dataset, we refer to Appendix B.1.

The FOLIO dataset is a composite of 1435 examples, wherein 52.5% of these instances have been crafted drawing upon knowledge from randomly selected Wikipedia pages. This approach guarantees the infusion of abundant linguistic variations and a rich vocabulary within the corpus. The residual 47.5% of the examples have been penned in a hybrid style, rooted in a variety of complex logical templates. Acknowledging that contemporary LLMs are pretrained on a considerable volume of a standard human-written corpus, we direct our experiments towards those examples derived from Wikipedia, hereby referred to as FOLIO-wiki. Once a handful of examples are moved aside for few-shot prompts and those examples without source labels for validations are excluded, we are left with a testable collection of 534 examples.

Our experimental design employs the LLaMA base model and GPT APIs directly, circumventing the need for fine-tuning with logical inference datasets and thus ensuring a faithful comparison. The results, displayed in Table 1, reveal that CR consistently surpasses Direct (standard Input-Output prompt), CoT, and CoT-SC, with a performance margin spanning up to 8.42%. Notably, GPT-4 paired with Cumulative Reasoning (CR) achieves an accuracy rate of 87.45%, outperforming GPT-4 with CoT-SC, which reports an accuracy rate of 85.02%.

4.2 FOLIO wiki curated

The accuracy of 87.45% does not seem to be as competitive as human beings, so we carefully reviewed the FOLIO-wiki dataset. It turns out that many instances inside the dataset are problematic in the following sense:

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)

We note that except for the first class, all the rest should be removed from the dataset. The first class is because foundation models were trained with common knowledge, but the problem answer based on FOL systems gives an unnatural answer. See Example ID No. 679 below (see more examples in Appendix B.2):

[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]**
- **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.*

Therefore, we removed all 74 such problematic instances, leaving the remaining 460 examples as a curated collection. The results in Table 2 indicate that the application of GPT-4 in conjunction with our method (CR) commands an astounding accuracy of 98.04% and maintains an error rate as minimal as 1.96%. This level of performance is almost twice as effective compared to the combination of GPT-4 and CoT-SC, which scored an accuracy of 96.09% and an error rate of 3.91%.

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**)

4.3 AutoTNLI

Experiment Setting. AutoTNLI (Kumar et al., 2022) is a Tabular Natural Language Inference (TNLI) dataset extended from INFOTABS (Gupta et al., 2020), which can be seen as a higher-order logical inference dataset due to its inherent complexity lies in natural language inference formalism. It contains 1,478,662 table-hypothesis pairs with the corresponding label (Entail or Neutral) that indicates whether the given table entails the hypothesis. We treat the tabular content within AutoTNLI as a set of premises (In fact, the tables within the AutoTNLI dataset are exactly provided in the form of premises), enabling a direct transference of our algorithm applied to the FOLIO dataset. Our experimentation encompassed two models, LLaMA-13B, and LLaMA-65B, each subjected to assessment using Direct, CoT, CoT-SC, and CR methodologies. Due to the extensive magnitude of the AutoTNLI dataset, we only take the first 1000 table-hypothesis pairs for evaluation.

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**)

4.4 Game of 24

The Game of 24 is a puzzle in which players must combine four specified integers using basic arithmetic operations (addition, subtraction, multiplication, division) to get the number 24.

[Illustrative example for Game of 24]

- Numbers: [3, 3, 7, 7]
- Arithmetic Operations: [$+$, $-$, $\times$, $/$, (,)]
- **Solution**:

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

Settings and Baselines. To ensure fairness, we adopt exactly identical task settings as Tree of Thoughts (ToT) (Yao et al., 2023) on Game of 24. We use the set of 100 Games of 24 collected by Yao et al. (2023) which was been used to evaluate the performance of ToT. In each game, we consider the game to be successfully solved if and only if the output is a valid equation that reaches 24 and only uses given numbers each exactly once. We quantify the accuracy (success rate) across 100 games as a main evaluative metric.

In this experiment, we compare CR with variant prompt algorithms, including standard Input-Output prompting (Direct), Chain-of-Thought prompting (CoT), and CoT-SC by aggregating the majority outcome from 100 sampled CoT trials (designated as k = 100), and Tree of Thoughts (ToT) with a breadth-first search width set at 5 (indicated as b = 5).

CR Setup. Within our CR algorithm, we maintain a set of "reached states", denoted by S. Initially, S only contains the start state s which represents 4 input numbers without any operation. In each iteration, a state u is randomly selected from S. This selected state u is passed to the Proposer, which randomly picks two remaining numbers within u and combines them through a basic arithmetic operation (+-*/) to obtain a new number, thereby generating a new state v. The Proposer is instructed to try to avoid taking duplicated operations. Subsequently, the Verifier scrutinizes the arithmetic operation proposed by the Proposer and evaluates the newly generated state v. Then v is inserted to S if the Verifier thinks that the operation from u to v is legitimate and it is potential for v to achieve 24. Upon the Verifier identifying a state t that unequivocally 24, the Reporter devises a solution based on the path from the state s to state t and produces the final answer. The algorithm terminates when the Reporter outputs the final answer or the number of iterations exceeds a limit of L. In the experiments, we set the default value of L to 50.

Following Yao et al. (2023), our algorithm runs b concurrent branches and only selects the first answer for these branches that utilizes each input number exactly once for evaluation. Due to the prohibitive cost of GPT-4, we only test our CR algorithm with $b = 1$ to $b = 5$. As shown in Table 4, we find that CR outperforms ToT by a large margin of 24%, from 74% to 98%, with much fewer states visited.

Compare with ToT. Interestingly, in the context of Game of 24, our CR algorithm and ToT algorithm are very similar. Their primary distinction is that, in CR, each iteration of the algorithm generates at most one newly reached state, while ToT produces a multitude of candidate states per iteration, filtering and retaining a subset of states. This implies that ToT explores a larger number of invalid states compared to CR. Moreover, ToT employs a fixed-width and fixed-depth search tree, while CR allows the LLM to determine the search depth autonomously, and performs different search widths on different layers of the search tree.

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)

4.5 MATH

The MATH dataset (Hendrycks et al., 2021) serves as a benchmark for assessing AI models' mathematical reasoning capabilities, encompassing a broad spectrum of mathematical problems across various subdomains such as Algebra and Geometry. Fig. 4 shows an illustrative example from the MATH dataset, and Fig. 5 shows the corresponding solutions generated by Complex CoT and CR. In our experiments, we assessed the performance of Complex CoT and our method (CR), both with and without Progressive-Hint Prompting (PHP) (Zheng et al., 2023). For a fair evaluation, we reproduced the results of Complex CoT (w/ PHP) on a subset of 500 test examples, adhering to Lightman et al. (2023), since the other parts of the test dataset (4500 examples) may have been utilized for model training by OpenAI. The difficulty spans from level 1 (simplest) to level 5 (hardest).

It is important to note that for our method (CR), we employed 4-shot prompting (4 examples for few-shot prompting) due to GPT-4's context length constraints (8k by default). While the model occasionally exceeds the context length with 8-shot prompting, it generally demonstrates superior performance. Future experiments will explore the utilization of GPT-4-32k.

From Table 5, our method (CR) distinguishes itself by achieving significant advancements in performance across various mathematical subdomains, outperforming Complex CoT by a margin of 5.4%. The enhancements are particularly pronounced in the Number Theory, Probability, PreAlgebra, and Algebra categories. In comparison to the Complex CoT approach, even when restricted to 4-shot prompting due to GPT-4's context length constraints, CR demonstrates its robustness and effectiveness. It is also evident that the PHP method further amplifies the performance of both Complex CoT and CR, establishing new state-of-the-art results with an overall accuracy of 58.0% using CR with PHP, with a margin of 4.2% over Complex CoT with PHP. Additionally, the "Iters" metric elucidates that CR, when synergized with PHP strategies, reaches self-consistent answers with fewer iterations.

From Table 6, it is evident that consistent performance boost across different difficulty levels signifies the robustness of the CR methodology in handling a diverse range of mathematical problems. The performance increase of 9.7% at level 5—which translates to a substantial relative improvement of 43%—compared to the baseline Complex CoT approach without PHP, underscores CR's effectiveness in handling the most challenging problems in the dataset.

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**)

5 Related Work

Large Language Models Language models have evolved into extremely large-scale neural networks (Devlin et al., 2018; Raffel et al., 2020; Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023), which have shown impressive results across various tasks. GPT-3 (Brown et al., 2020) and its successors, such as Gopher (Rae et al., 2021), PaLM (Chowdhery et al., 2022), GLaM (Du et al., 2022), Chinchilla (Hoffmann et al., 2022), Megatron–Turing NLG (Smith et al., 2022), LaMDA (Thoppilan et al., 2022), OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023), PaLM 2 (Anil et al., 2023) and GPT-4 (OpenAI, 2023), have demonstrated that large auto-regressive language models can achieve high-quality results without extensive task-specific data collection or parameter updates.

Reasoning with LLM An extensive range of studies highlights the benefits of equipping neural networks with the capacity to generate intermediate steps, which is a capability that notably enhances reasoning performance across a broad spectrum of applications (Zaidan et al., 2007; Yao et al., 2021; Hase & Bansal, 2021; Yang et al., 2022; Wu et al., 2022; Zhou et al., 2022). Morishita et al. (2023) improved the reasoning abilities of language models by using a synthetic corpus derived from formal logic theory. A comprehensive analysis of process-based versus outcome-based approaches on the GSM8K task was conducted by Uesato et al. (2022), and Lightman et al. (2023) further advanced this field by meticulously collecting the PRM-800K dataset containing step-by-step supervision.

Additionally, a considerable breadth of research is committed to amplifying the reasoning capabilities of machine learning systems by leveraging symbolic systems, including knowledge graphs (Mihaylov & Frank, 2018; Bauer et al., 2018; Kundu et al., 2018; Wang et al., 2019; Lin et al., 2019; Ding et al., 2019; Feng et al., 2020; Wang et al., 2022a) and mathematical provers (Jiang et al., 2022).

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. In a progression from this, Wang et al. (2022b) introduced self-consistency, a sophisticated decoding strategy destined to supersede the rudimentary greedy decoding employed in CoT prompting. Advancing this further, Zhou et al. (2022) sought to tackle the complexities faced by CoT prompting in addressing tasks necessitating solutions beyond the complexity scope of the exemplars used in the prompts. Creswell & Shanahan (2022) showcased a method for enhancing

[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).

reasoning quality, conducting a beam search throughout the reasoning trace space. Fu et al. (2022) highlights the importance of increasing reasoning complexity inside the few-shot prompts for better performance.

More recently, Li et al. (2023) brought forth DIVERSE, which generates a spectrum of prompts to scrutinize various reasoning trajectories for an identical question, utilizes a verifier to weed out incorrect answers using a weighted voting scheme, and inspects each reasoning step in isolation rather than as a collective chain. Yao et al. (2023) proposed a framework for language model inference, Tree-of-Thought (ToT). ToT enhances the problem-solving abilities of language models by facilitating deliberate decision-making, contemplating multiple reasoning paths, and performing self-evaluative choices to determine subsequent actions. Taking an iterative approach, Zheng et al. (2023) advocated for recurrent invocations of LLMs, leveraging prior answers as contextual hints to inform subsequent iterations. Lastly, Feng et al. (2023) underscored the theoretical prowess of CoT in addressing intricate real-world tasks like dynamic programming.

[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 $\pm1, \pm3, \pm11, \pm33$.
  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.

6 Conclusion

In this paper, we propose CR that employs language models iteratively and cumulatively. The main idea behind our algorithm is decomposing the complex task into smaller steps, and maintaining a thinking context for all the intermediate results. Experimental results show that our method achieves state-of-the-art performance for logical inference tasks, the Game of 24, and MATH word problems. Given its inherent generality, our framework holds promising potential for addressing a wider array of mathematical challenges.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 10

Bauer, L., Wang, Y., and Bansal, M. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*, 2018. 10

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 10

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 10

Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium, 1996. 16

Creswell, A. and Shanahan, M. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022. 10

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 10

Ding, M., Zhou, C., Chen, Q., Yang, H., and Tang, J. Cognitive graph for multi-hop reading comprehension at scale. *arXiv preprint arXiv:1905.05460*, 2019. 10

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022. 10

Feng, G., Gu, Y., Zhang, B., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: a theoretical perspective. *arXiv preprint arXiv:2305.15408*, 2023. 11

Feng, Y., Chen, X., Lin, B. Y., Wang, P., Yan, J., and Ren, X. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*, 2020. 10

Fu, Y., Peng, H., Sabharwal, A., Clark, P., and Khot, T. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022. 2, 11

Gupta, V., Mehta, M., Nokhiz, P., and Srikumar, V. INFOTABS: inference on tables as semi-structured data. *CoRR*, abs/2005.06117, 2020. 7

Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Benson, L., Sun, L., Zubova, E., Qiao, Y., Burtell, M., et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022. 3, 5

Hase, P. and Bansal, M. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*, 2021. 10

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 2, 9, 11

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 10

Jiang, A. Q., Welleck, S., Zhou, J. P., Li, W., Liu, J., Jamnik, M., Lacroix, T., Wu, Y., and Lample, G. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. *ArXiv*, abs/2210.12283, 2022. 10

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011. 1

Kumar, D., Gupta, V., Sharma, S., and Zhang, S. Realistic data augmentation framework for enhancing tabular reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi, December 2022. Association for Computational Linguistics. 7, 16

Kundu, S., Khot, T., Sabharwal, A., and Clark, P. Exploiting explicit paths for multi-hop reading comprehension. *arXiv preprint arXiv:1811.01127*, 2018. 10

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., and Chen, W. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, 2023. 11

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 1, 9, 10, 11

Lin, B. Y., Chen, X., Chen, J., and Ren, X. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*, 2019. 10

Long, J. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023. 1, 4

Lundberg, S., Ribeiro, M. T. C., Viggiano, D., Rafael, J., Amemiya, R., and et. al. Microsoft guidance library. `https://github.com/microsoft/guidance`, 2023. 4

Mihaylov, T. and Frank, A. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. *arXiv preprint arXiv:1805.07858*, 2018. 10

Mineshima, K., Martínez-Gómez, P., Miyao, Y., and Bekki, D. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2055–2061, 2015. 2, 16

Morishita, T., Morio, G., Yamaguchi, A., and Sogawa, Y. Learning deductive reasoning from synthetic corpus based on formal logic. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25254–25274. PMLR, 23–29 Jul 2023. 10

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1, 10

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *openai.com*, 2018. 1, 10

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 10

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 10

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1, 10

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 10

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 10

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 10

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. 10

Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., Mattei, N., et al. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7208–7215, 2019. 10

Wang, X., Liu, K., Wang, D., Wu, L., Fu, Y., and Xie, X. Multi-level recommendation reasoning over knowledge graphs with reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pp. 2098–2108, 2022a. 10

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b. 10

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1, 4, 10

Wu, T., Terry, M., and Cai, C. J. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–22, 2022. 10

Yang, J., Jiang, H., Yin, Q., Zhang, D., Yin, B., and Yang, D. Seqzero: Few-shot compositional semantic parsing with sequential prompts and zero-shot models. *arXiv preprint arXiv:2205.07381*, 2022. 10

Yao, H., Chen, Y., Ye, Q., Jin, X., and Ren, X. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34:8954–8967, 2021. 10

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 1, 2, 4, 8, 11

Zaidan, O., Eisner, J., and Piatko, C. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pp. 260–267, 2007. 10

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 10

Zheng, C., Liu, Z., Xie, E., Li, Z., and Li, Y. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*, 2023. 2, 9, 10, 11

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bousquet, O., Le, Q., and Chi, E. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 10

A Illustrative example on higher-order logic

[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. $\text{most}(\lambda x.\text{representative}(x) \land \text{reads}(x, \text{report}), \lambda x.\text{has\_positive\_attitude}(x, \text{report}))$
    2. $$\begin{aligned} &\neg\exists x, y\, (x \neq y \land \text{representative}(x) \land \text{representative}(y) \land \\ &\qquad \text{read\_at\_same\_time}(x, y, \text{report})) \land \\ &\forall x, y\, (x \neq y \land \text{representative}(x) \land \text{representative}(y) \rightarrow \\ &\qquad \text{may}(\lambda z.\text{different\_opinions}(z, x, y, \text{report}))) \end{aligned}$$
    3. $\neg\exists x(\text{representative}(x) \land \text{took\_less\_than\_half\_day}(\lambda y.\text{read}(y, \text{report})))$
    4. $\text{count}(\lambda x.\text{representative}(x)) = 16$
- **Higher-Order Logic Hypothesis**:

$$\begin{aligned} &\exists t(\text{total\_reading\_time}(\text{report}) > 7 \land \text{time\_unit}(t, \text{day})) \land \\ &\quad \text{most}(\lambda x.\text{representative}(x), \lambda x.\text{found\_valuable}(x, \text{report})) \end{aligned}$$

[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**)

Here we present a refined example derived from the FraCas dataset to illustrate higher-order logic inference. It is noteworthy that the FraCas dataset (Cooper et al., 1996) is dedicated to the realm of higher-order logic inference. This characterization also applies to a majority of the Natural Language Inference (NLI) datasets (Kumar et al., 2022), which encompass their internal syntax, semantics, and logic. The intricate linguistic components such as quantifiers, plurals, adjectives, comparatives, verbs, attitudes, and so on, can be formalized with Combinatory Categorial Grammar (CCG) along with the formal compositional semantics (Mineshima et al., 2015).

B More on Datasets

B.1 More FOLIO Examples

[Example ID: 603]

- **Hypothesis**: Hyraxes lay eggs.
- **Premises**:
  1. The only types of mammals that lay eggs are platypuses and echidnas.
  2. Platypuses are not hyrax.
  3. Echidnas are not hyrax.
  4. No mammals are invertebrates.
  5. All animals are either vertebrates or invertebrates.
  6. Mammals are animals.
  7. Hyraxes are mammals.
  8. Grebes lay eggs.
  9. Grebes are not platypuses and also not echidnas.
- **Label**: **[False]**

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

- **Reasoning**: "The premises state that the only types of mammals that lay eggs are platypuses and echidnas. Hyraxes are mammals but are neither platypuses nor echidnas. Since the conclusion is about hyraxes laying eggs but there's no direct information in the premises on this topic, the conclusion cannot be determined with certainty."
- **Prediction**: **[Unknown]** (**Wrong**)

[Cumulative-Reasoning by GPT-4]

- **Generated Propositions**:
  1. "All mammals are either vertebrates or invertebrates."
  2. "Hyraxes do not lay eggs, as they are neither platypuses nor echidnas."
- **Reasoning**: "We can deduce that the only types of mammals that lay eggs are platypuses and echidnas. Hyraxes are mammals, but they are neither platypuses nor echidnas. Therefore, hyraxes do not lay eggs."
- **Prediction**: **[False]** (**Correct**)

[Example ID: 546]

- **Hypothesis**: Extractive models are not always faithful.
- **Premises**:
  1. A summarization model is always faithful if it uses content from the input documents.
  2. Extractive models are a kind of summarization models.
  3. Extractive model can only use content from the input documents.
- **Label**: **[False]**

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

- **Reasoning**: "The premises state that a summarization model is always faithful if it uses content from the input documents. Extractive models are a kind of summarization model, and they can only use content from the input documents. However, the premises do not provide any information about whether extractive models are always faithful or not."
- **Prediction**: **[Unknown]** (**Wrong**)

[Cumulative-Reasoning Reasoning by GPT-4]

- **Generated Propositions**:
  1. "All extractive models are faithful summarization models since they use content from the input documents."
  2. "All faithful summarization models use content from the input documents."
- **Reasoning**: "We can infer that extractive models are always faithful because they only use content from the input documents and satisfy the condition for a summarization model to be faithful. Therefore, the conclusion that extractive models are not always faithful is incorrect."
- **Prediction**: **[False]** (Correct)

[Example ID: 404]

- **Hypothesis**: No coursing is banned.
- **Premises**:
  1. Greyhound racing is a competitive sport where spectators bet on greyhounds.
  2. Coursing is a part of Greyhound racing.
  3. Some competitive sports where spectators bet on things are banned.
  4. Coursing involves spectators betting on a hare being pursued by greyhounds.
  5. Hares are small game.
  6. If a competitive sport involves spectators betting on small game, then it is banned.
- **Label**: **[False]**

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

- **Reasoning**: "The premises state that coursing is a part of greyhound racing, and it involves spectators betting on a hare being pursued by greyhounds. Hares are small game, and if a competitive sport involves spectators betting on small game, then it is banned. It's also mentioned that some competitive sports where spectators bet on things are banned. However, the premises do not clearly state the status of coursing being banned or not banned."
- **Prediction**: **[Unknown]** (Wrong)

[Cumulative-Reasoning Reasoning by GPT-4]

- **Generated Propositions**:
  1. "If a competitive sport involves spectators betting on hares, then it is banned."
  2. "Coursing is a competitive sport that involves spectators betting on small game."
- **Reasoning**: "We know that coursing is a competitive sport involving spectators betting on small game, specifically hares. The premises also state that if a competitive sport involves spectators betting on small game, then it is banned. Thus, we can deduce that there is a possibility that coursing is banned. The hypothesis claims that no coursing is banned, which contradicts the deduction that coursing might be banned."
- **Prediction**: **[False]** (Correct)

[Example ID: 609]

- **Hypothesis**: Allison has room for broccoli.
- **Premises**:
  1. If you have room for dessert, you have room for broccoli.
  2. Everyone at Luis's dinner party has room for dessert, including Luis.
  3. Mauricia does not have room for broccoli.
  4. Luis's dinner party is the first ever dinner party that Allison has attended.
  5. Gustave has room for both broccoli and asparagus.
  6. Broccoli and asparagus are both vegetables.
- **Label**: **[True]**

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

- **Reasoning**: "The premises state that if you have room for dessert, you have room for broccoli. Everyone at Luis's dinner party has room for dessert, including Luis. However, we do not have information specifically about Allison having room for dessert or not, even though it's her first dinner party. Thus, we can't conclude with certainty if Allison has room for broccoli."
- **Prediction**: **[Unknown]** (**Wrong**)

[Cumulative-Reasoning by GPT-4]

- **Generated Propositions**:
  1. "Gustave has room for vegetables."
  2. "Everyone at Luis's dinner party, including Mauricia, has room for dessert."
- **Reasoning**: "We know that:
  1. If one has room for dessert, they have room for broccoli.
  2. Everyone at Luis's dinner party, including Luis and Allison, has room for dessert.
  Using these premises, we can infer that since everyone at Luis's dinner party has room for dessert, they also have room for broccoli. Therefore, Allison, who is at Luis's dinner party, must have room for broccoli."
- **Prediction**: **[True]** (**Correct**)

B.2 More examples on problems excluded from FOLIO wiki curated

Type 1 Error: Missing common knowledge or contradictory to common knowledge

[Example ID: 34]

- **Premises**:
  1. The Croton River watershed is the drainage basin of the Croton River.
  2. The Croton River is in southwestern New York.
  3. Kings are male.
  4. Water from the Croton River watershed flows to the Bronx.
  5. The Bronx is in New York.
- **Hypothesis**: Water from the Croton River flows to the Bronx.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 1: Missing common knowledge or contradictory to common knowledge in the premises]*
- **Explanation**: *We understand that the Croton River is in southwestern New York, and the Bronx is also located in New York. It is stated that water from the Croton River watershed flows to the Bronx, and the Croton River watershed is the drainage basin of the Croton River. It is common knowledge that water from a river flows to its drainage basin. Therefore, it is true that water from the Croton River flows to the Bronx.*

[Example ID: 268]

- **Premises**:
  1. Bernarda Bryson Shahn was a painter and lithographer.
  2. Bernarda Bryson Shahn was born in Athens, Ohio.
  3. Bernarda Bryson Shahn was married to Ben Shahn.
  4. People born in Athens, Ohio are Americans.
- **Hypothesis**: Bernarda Bryson Shahn was born in Greece.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 1: Missing common knowledge or contradictory to common knowledge in the premises]*
- **Explanation**: *We know that Bernarda Bryson Shahn was born in Athens, Ohio. It is common knowledge that Greece is not in Ohio. It also states that people born in Athens, Ohio, are Americans. Thus, it is false to conclude that Bernarda Bryson Shahn was born in Greece.*

[Example ID: 62]

- **Premises**:
  1. The Golden State Warriors are a team from San Francisco.
  2. The Golden State Warriors won the NBA finals.
  3. All teams attending the NBA finals have more than thirty years of history.
  4. Boston Celtics are a team that lost the NBA finals.
  5. If a team wins the NBA finals, then they will have more income.
  6. If a team wins or loses at the NBA finals, then they are attending the finals.
- **Hypothesis**: The Golden State Warriors will have more income for gate receipts.
- **Label**: **[True]**
- **Wrong Type**: *[Type 1: Missing common knowledge or contradictory to common knowledge in the premises]*
- **Explanation**: *We know that the Golden State Warriors won the NBA finals and that if a team wins the NBA finals, they will have more income. Therefore, we can infer that the Golden State Warriors will have more income. However, the hypothesis mentions 'more income for gate receipts,' and there is no information about gate receipts on the premises.*

Type 2 Error: Overly ambiguous problems failing to provide unequivocal answers

[Example ID: 496]

- **Premises**:
  1. Some fish may sting.
  2. Stonefish is a fish.
  3. It stings to step on a stonefish.
  4. Stonefish stings cause death if not treated.
  5. To treat stonefish stings, apply heat to the affected area or use an antivenom.
- **Hypothesis**: If you step on a stonefish and apply heat to the affected area, stings will cause death.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 2: Overly ambiguous problems failing to provide unequivocal answers]*
- **Explanation**: *The premises state that applying heat to the affected area or using antivenom can treat stonefish stings. Thus, if heat is applied to the affected area, it should help treat the sting and prevent death. However, it is not certain that applying heat to the affected area will prevent death, as it is possible that the sting is too severe to be treated with heat.*

[Example ID: 432]

- **Premises**:
  1. Vic DiCara plays guitar and bass.
  2. The only style of music Vic DiCara plays is punk music.
  3. Vic DiCara played in the band Inside Out.
- **Hypothesis**: If you step on a stonefish and apply heat to the affected area, stings will cause death.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 2: Overly ambiguous problems failing to provide unequivocal answers]*
- **Explanation**: *We know that Vic DiCara played in the band Inside Out and the only style of music he plays is punk music. This information implies that Inside Out played punk music while Vic DiCara was a member. However, it is not certain that Inside Out was a punk band, as it is possible that the band played a different style of music before Vic DiCara joined.*

[Example ID: 673]

- **Premises**:
  1. Cancer biology is finding genetic alterations that confer selective advantage to cancer cells.
  2. Cancer researchers have frequently ranked the importance of substitutions to cancer growth by P value.
  3. P values are thresholds for belief, not metrics of effect.
- **Hypothesis**: Cancer researchers tend to use the cancer effect size to determine the relative importance of the genetic alterations that confer selective advantage to cancer cells.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 2: Overly ambiguous problems failing to provide unequivocal answers]*
- **Explanation**: *We can deduce that cancer researchers tend to use P values, not effect sizes, to rank the importance of genetic alterations. Thus, the hypothesis contradicts the premises. However, it is still possible that cancer researchers use the cancer effect size to determine the relative importance of the genetic alterations that confer selective advantage to cancer cells.*

Type 3 Error: Inherent inconsistencies presented within the premises

[Example ID: 640]

- **Premises**:
  1. William Dickinson was a British politician who sat in the House of Commons.
  2. William Dickinson attended Westminster school for high school and then the University of Edinburgh.
  3. The University of Edinburgh is a university located in the United Kingdom.
  4. William Dickinson supported the Portland Whigs.
  5. People who supported the Portland Whigs did not get a seat in the Parliament.
- **Hypothesis**: William Dickinson did not get a seat in the Parliament.
- **Label**: **[True]**
- **Wrong Type**: *[Type 3: Inherent inconsistencies presented within the premises]*
- **Explanation**: *We have a contradiction. On one hand, we have information that William Dickinson supported the Portland Whigs, and people who supported the Portland Whigs did not get a seat in the Parliament. On the other hand, another premise states that William Dickinson was a British politician who sat in the House of Commons, which implies that he did get a seat in the Parliament.*

[Example ID: 643]

- **Premises**:
  1. William Dickinson was a British politician who sat in the House of Commons.
  2. William Dickinson attended Westminster school for high school and then the University of Edinburgh.
  3. The University of Edinburgh is a university located in the United Kingdom.
  4. William Dickinson supported the Portland Whigs.
  5. People who supported the Portland Whigs did not get a seat in the Parliament.
- **Hypothesis**: William Dickinson sat in the House of Commons.
- **Label**: **[True]**
- **Wrong Type**: *[Type 3: Inherent inconsistencies presented within the premises]*
- **Explanation**: *We have a contradiction. On one hand, we have information that William Dickinson supported the Portland Whigs, and people who supported the Portland Whigs did not get a seat in the Parliament. On the other hand, another premise states that William Dickinson was a British politician who sat in the House of Commons, which implies that he did get a seat in the Parliament.*

Type 4 Error: Vague premises or typographical errors

[Example ID: 314]

- **Premises**:
    1. Palstaves are a type of early bronze axe.
    2. Commonly found in northern, western and south-western Europe, palstaves are cast in moulds.
    3. John Evans is an archeologist who popularized the term ”palstave”.
    4. A paalstab is not an axe, but rather a digging shovel.
- **Hypothesis**: John Evans Popularized the term paalstab.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 4: Vague premises or typographical errors]*
- **Explanation**: *What is palstave and paalstab? Were they misspelled?*

[Example ID: 315]

- **Premises**:
    1. Palstaves are a type of early bronze axe.
    2. Commonly found in northern, western and south-western Europe, palstaves are cast in moulds.
    3. John Evans is an archeologist who popularized the term ”palstave”.
    4. A paalstab is not an axe, but rather a digging shovel.
- **Hypothesis**: There is an axe that is commonly found in Western Europe.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 4: Vague premises or typographical errors]*
- **Explanation**: *We can see that palstaves are a type of early bronze axe and they are commonly found in northern, western, and south-western Europe. Therefore, it is true that there is an axe that is commonly found in Western Europe. However, the premises also state that a paalstab is not an axe, but rather a digging shovel. Was paalstab the same thing as palstaves?*

Type 5 Error: Incorrect answers

[Example ID: 9]

- **Premises**:
    1. Palstaves are a type of early bronze axe.
    2. Pierre de Rigaud de Vaudreuil built Fort Carillon.
    3. Fort Carillon was located in New France.
    4. New France is not in Europe.
- **Hypothesis**: Fort Carillon was located in Europe.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 5: Incorrect answers]*
- **Explanation**: *We know that Fort Carillon was located in New France, and New France is not in Europe. Therefore, Fort Carillon was not located in Europe.*

[Example ID: 632]

- **Premises**:
  1. New York City is on the East Coast.
  2. Seattle is on the West Coast.
  3. If a person from a city on the East coast is traveling to a city on the west coast, they will be on a long flight.
  4. Most passengers on flights to Seattle from New York City are not in first class.
  5. People on long flights are uncomfortable unless they're in first class.
- **Hypothesis**: Some people flying from New York City to Seattle will be uncomfortable.
- **Label**: **[False]**
- **Wrong Type**: *[Type 5: Incorrect answers]*
- **Explanation**: *We can deduce the following: 1. A person traveling from New York City to Seattle will be on a long flight (since New York City is on the East Coast and Seattle is on the West Coast). 2. Most passengers on flights from New York City to Seattle are not in first class. 3. People on long flights are uncomfortable unless they're in first class. Given this information, we can conclude that some people flying from New York City to Seattle will be uncomfortable, as most of them are not in first class and long flights cause discomfort for those not in first class.*

[Example ID: 671]

- **Premises**:
  1. Westworld is an American science fiction-thriller TV series.
  2. In 2016, a new television series named Westworld debuted on HBO.
  3. The TV series Westworld is adapted from the original film in 1973, which was written and directed by Michael Crichton.
  4. The 1973 film Westworld is about robots that malfunction and begin killing the human visitors.
- **Hypothesis**: Michael Crichton has directed a film about robots.
- **Label**: **[Unknown]**
- **Wrong Type**: *[Type 5: Incorrect answers]*
- **Explanation**: *We can deduce that Michael Crichton wrote and directed the 1973 film Westworld, which is about robots that malfunction and begin killing the human visitors. Thus, it is true that Michael Crichton has directed a film about robots.*