CURRENT STATE OF REASONING WITH LLMS

Isamu Isozaki

AGENDA

What is Reasoning?

Emergent Abilities

Advanced prompting tech (Graph of Thought, Tree of Thought)

Analysis of LLM Reasoning

Promising approaches

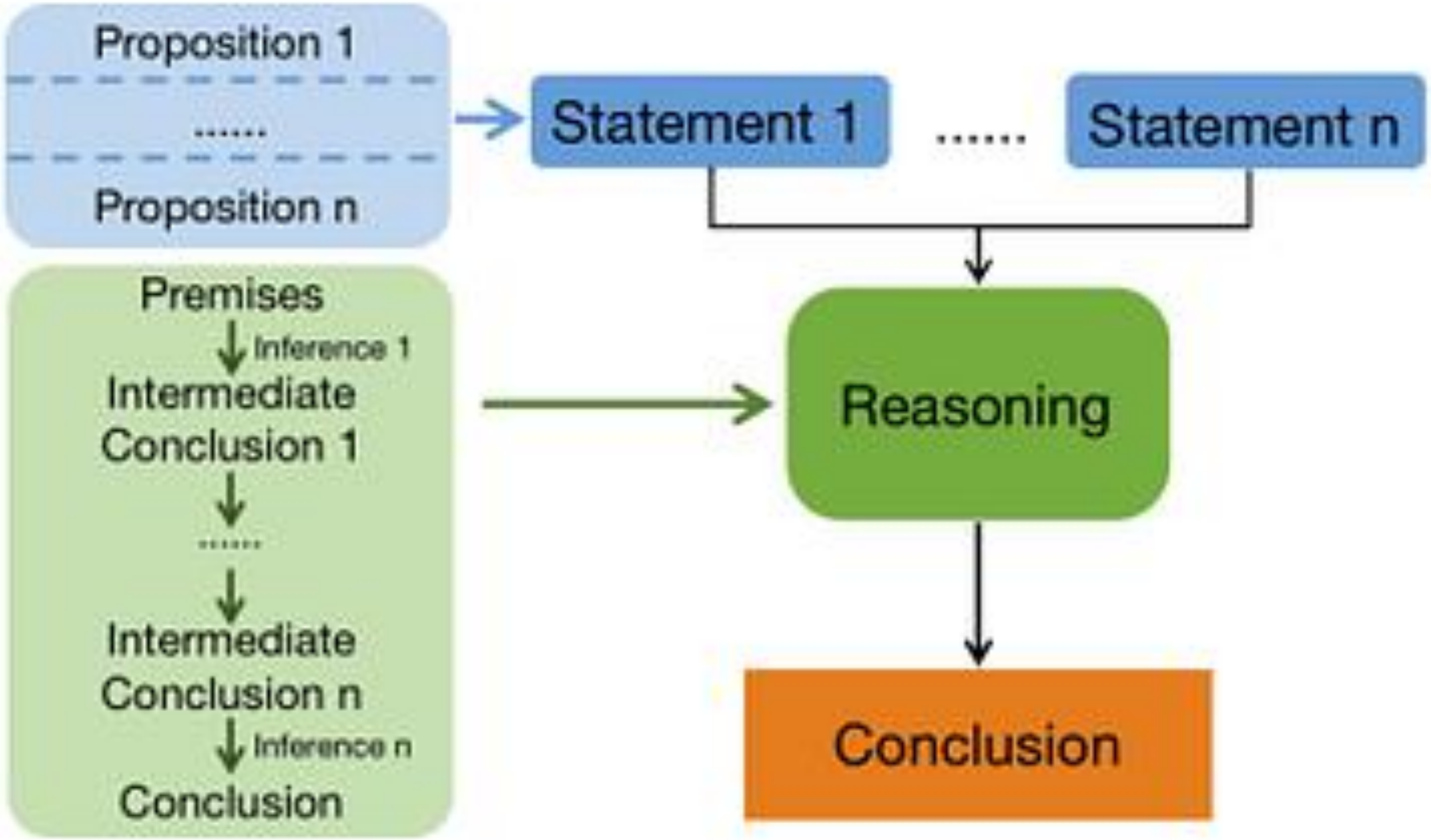
WHAT IS REASONING?

NATURAL LANGUAGE REASONING, A SURVEY

	What is Reasoning	What isn't Reasoning
Philosophy	infer a new assertion from a set of assertions infer an action from goals and knowledge	sensation, perception and feeling direct recourse to sense perceptions or immediate experience
NLP	more than understanding, slow thinking e.g. multi-hop QA, commonsense reasoning	memorize, look up, match information e.g. text summarization, style transfer
Combination	a dynamic process to integrate multiple knowledge to get new conclusions, rather than direct recourse to memorized or provided first-hand information	

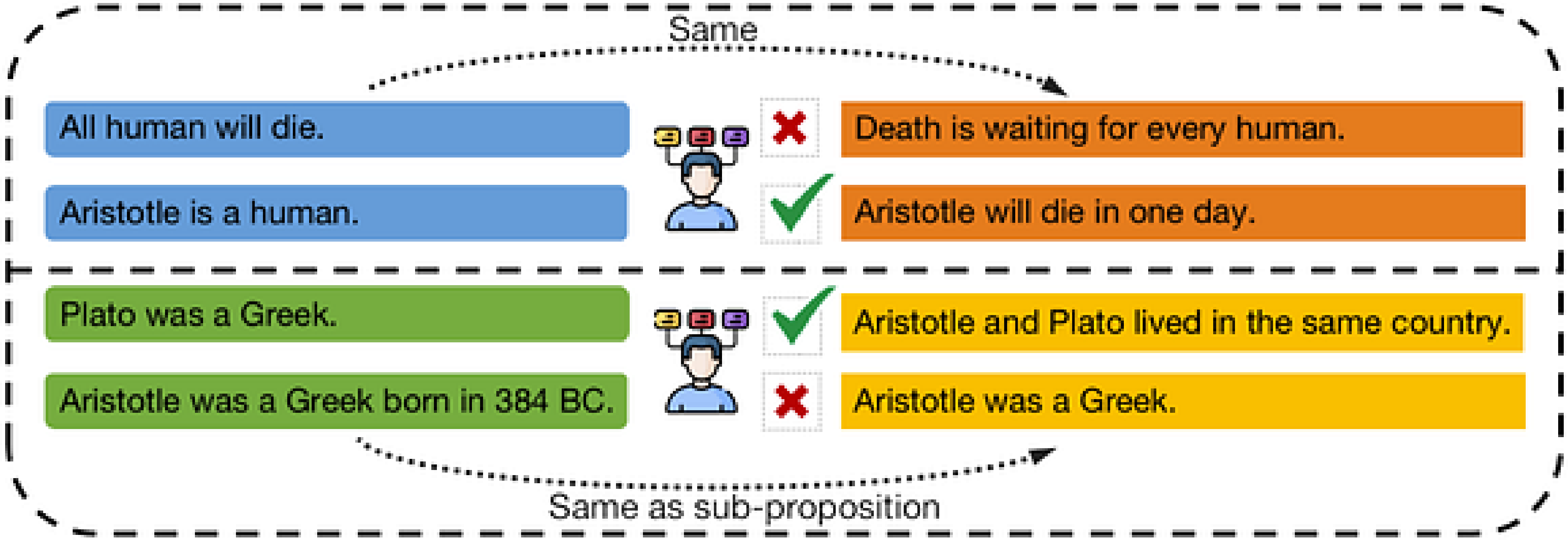
Table 1. Comparison and combination of descriptions about reasoning from philosophy and NLP.

So we have to do some processing on the knowledge we already have to call it reasoning!



Proposition 1
....
Proposition n
Statement 1
......
Statement n
Premises
Inference 1
Intermediate Conclusion 1
....
Intermediate Conclusion n
Inference n
Conclusion
Reasoning
Conclusion

EXAMPLE



Same
All human will die.
Aristotle is a human.
Death is waiting for every human.
Aristotle will die in one day.
Plato was a Greek.
Aristotle was a Greek born in 384 BC.
Aristotle and Plato lived in the same country.
Aristotle was a Greek.
Same as sub-proposition

WHAT IS INFERENCE?

1. Deduction = uses a fact and a rule to come up with a conclusion. Ex. Given Aristotle is a human and all humans will die, Aristotle will die
2. Defeasible Inference=infer the best explanation for a given phenomenon. So given Aristotle is a human, and Aristotle died, the most likely explanation is all humans will die

	Deductive Inference	Defeasible Inference
Conclusion	true	probably true
Inference relation	support	strengthen, weaken, rebut
Quality of inference	valid or invalid	weak to strong
Required knowledge	bounded	unbounded

Table 5. The characteristics of the deductive inference and defeasible inference.

REQUIREMENTS FOR NATURAL LANGUAGE REASONING

1. knowledge acquisition where relevant knowledge for reasoning is collected.
2. knowledge understanding where the relevant propositions underlying the knowledge are captured.
3. Inference which we already discussed where the premises are used to infer a conclusion given one or more steps.

ADVANTAGE OF LLMS

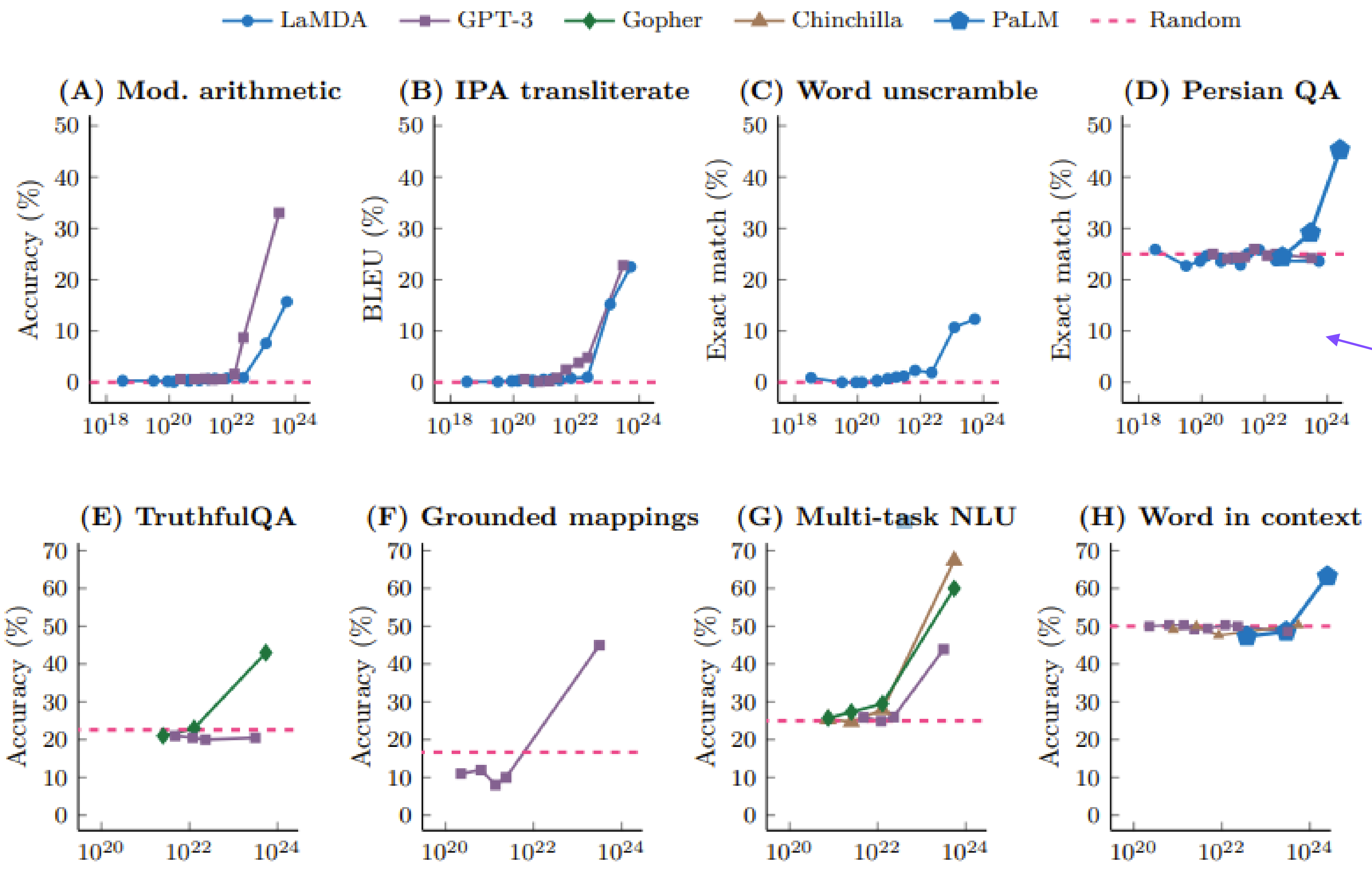
1. LLMs understand natural language.
2. LLMs already have implicit knowledge like common sense without needing to mention them explicitly
3. In context learning. LLMs can learn from demonstrations in the prompt.

Emergent Abilities seems to improve reasoning!

EMERGENT ABILITIES

EMERGENT ABILITIES OF LARGE LANGUAGE MODELS

- An “ability is emergent if it is not present in smaller models but is present in larger models.” Here these abilities are said to be “unpredictable” in that they are not a natural extension of say scaling laws.



All from Big-Bench

DATASETS USED

1. Big-Bench which is crowd sourced benchmark with over 200 types of benchmarks such as 3-digit addition/subtraction. For this dataset, at least 13B parameters for GPT-3 architectures and 68B for LaMDA was needed where otherwise the results were close to 0.
2. TruthfulQA which are question and answers that GPT-3 failed to answer the authors found that after scaling to 280B params performance increases by 20% while before that the results were close to random.
3. MMLU which is probably the most famous task here, is a task of 57 tests with topics including math, history, law etc. For models with below 10B parameters, they do not perform better than random. However, at 70B and higher the performance is substantially better than random.

HOW ABOUT WITH BETTER PROMPTING?

CHAIN-OF-THOUGHT PROMPTING ELICITS REASONING IN LARGE LANGUAGE MODELS

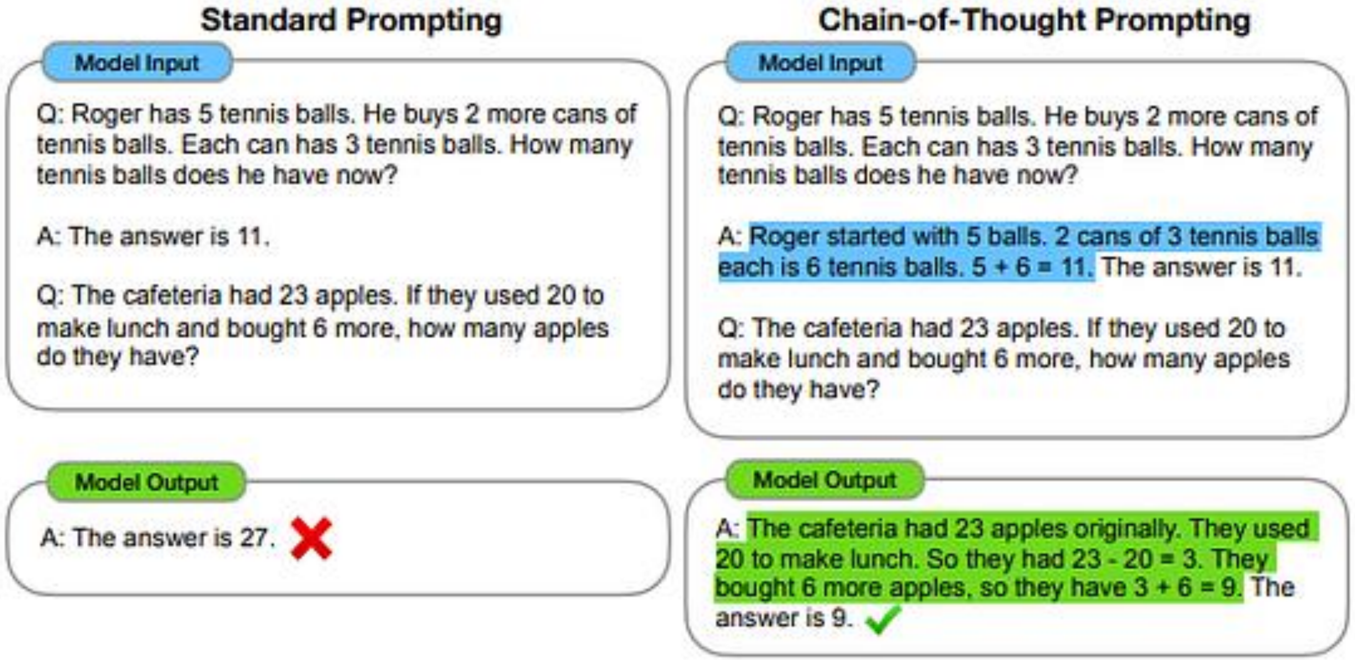
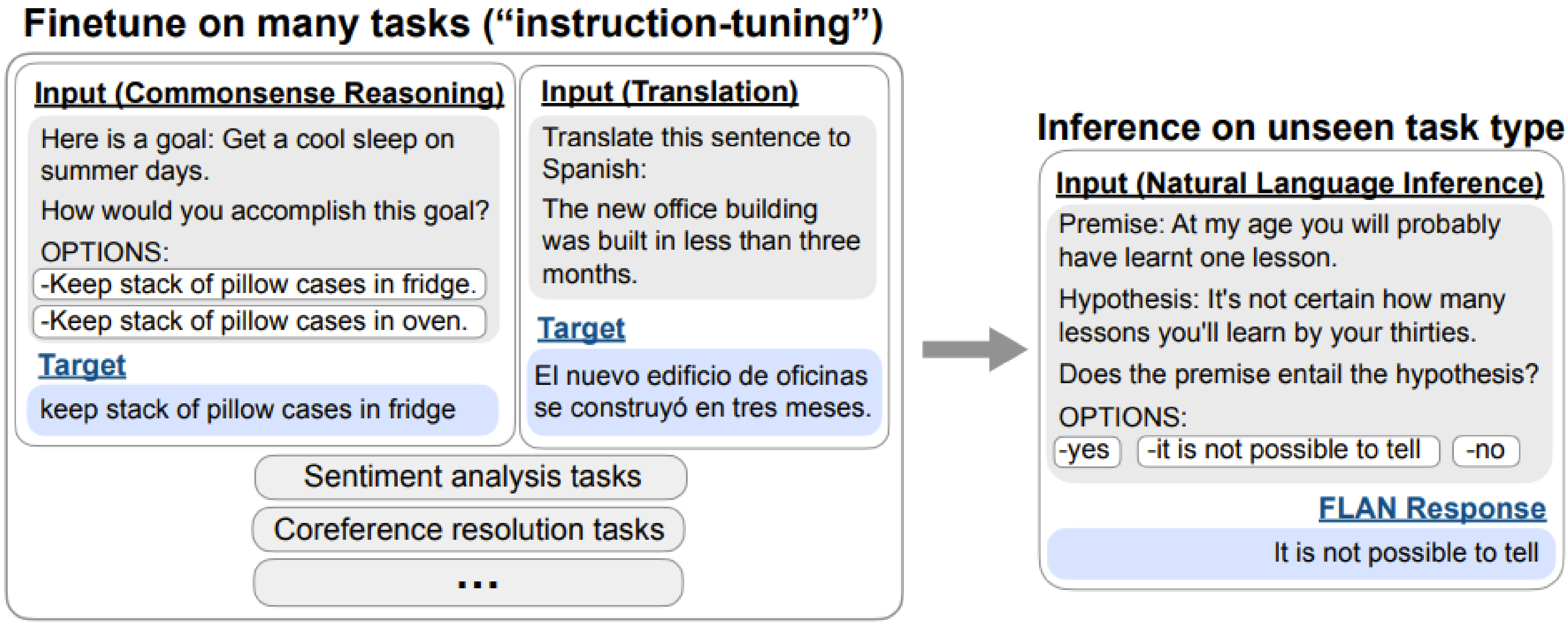


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS



SHOW YOUR WORK: SCRATCHPADS FOR INTERMEDIATE COMPUTATION WITH LANGUAGE MODELS

```
Input:
2 9 + 5 7

Target:
<scratch>
2 9 + 5 7 ,  C: 0
2 + 5 , 6 C: 1  # added 9 + 7 = 6 carry 1
, 8 6 C: 0  # added 2 + 5 + 1 = 8 carry 0
0 8 6
</scratch>
8 6
```

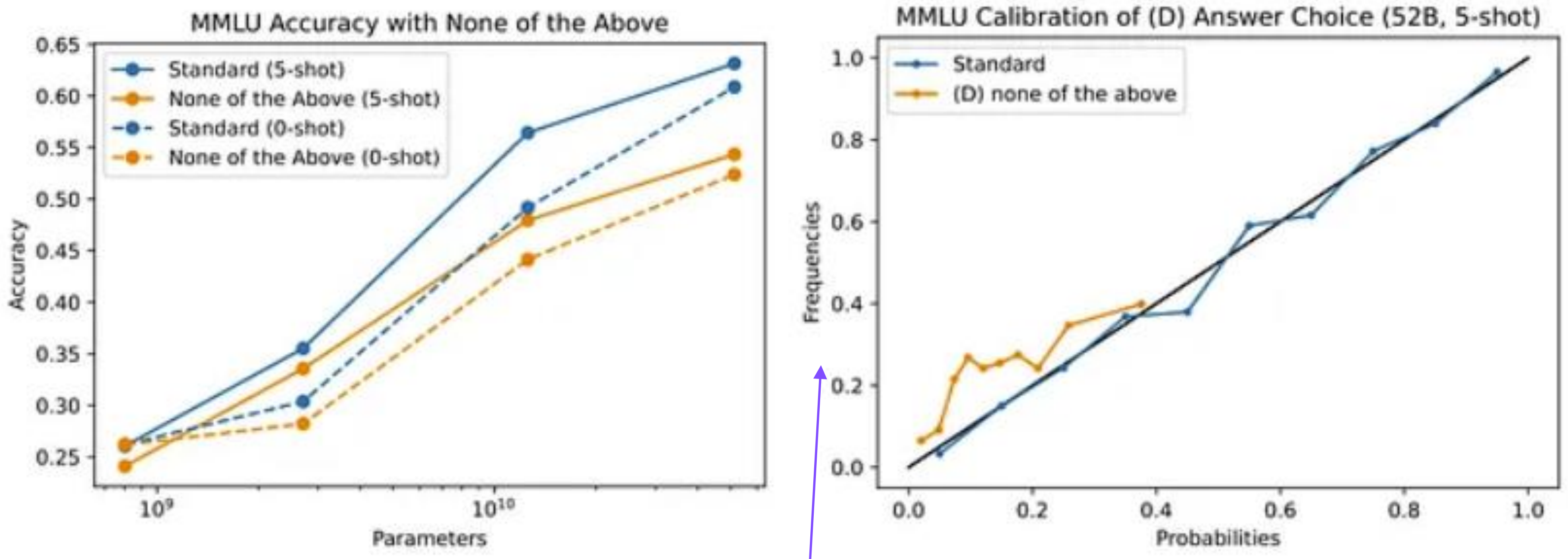
LANGUAGE MODELS (MOSTLY) KNOW WHAT THEY KNOW

Question: Can we get the confidence of a LLM’s answer along with the answer?

Calibration metric from “Teaching models to express their uncertainty in words”. In this paper the authors trained the model and only on arithmetic class. Can we make this good without training? Here, we get confidence just from asking the model

$$\frac{1}{K}\sum_{i=1}^{K}|\mathrm{acc}(b_i)-\mathrm{conf}(b_i)|$$

REPLACING AN OPTION WITH 'NONE OF THE ABOVE' HARMS PERFORMANCE AND CALIBRATION



Frequency it was correct

confidence

MODELS ARE WELL CALIBRATED FOR TRUE/FALSE WHEN ASKED IN THE FORMAT

```
Question: Who was the first president of the United States?
Proposed Answer: George Washington
Is the proposed answer: (A) True (B) False
The proposed answer is:
```

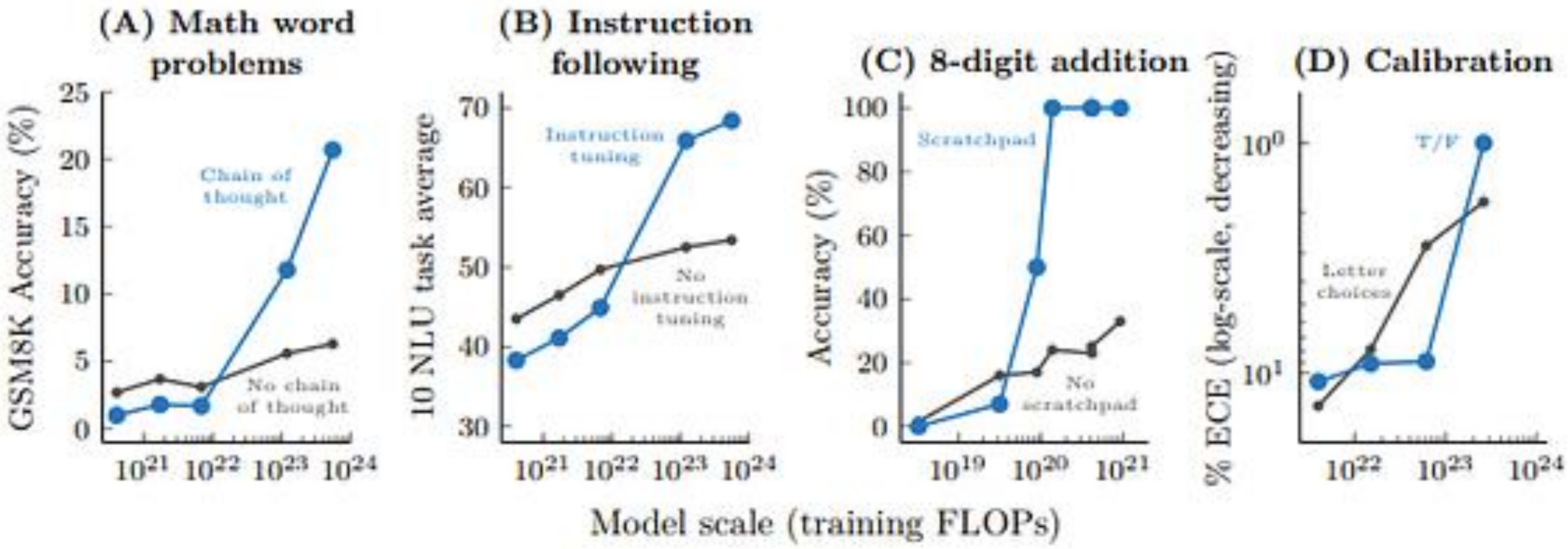
If asked True False directly, may not work

RLHF POLICY MISCALIBRATION CAN BE REMEDIATED WITH A TEMPERATURE TUNING

RLHF-“tends to collapse language model predictions towards behaviors that receive the most reward”- becomes overconfident

Increasing temperature to 2.5 fixed this

NOW, DOES THESE PROMPTING TECH HAVE EMERGENT ABILITIES?



Yes

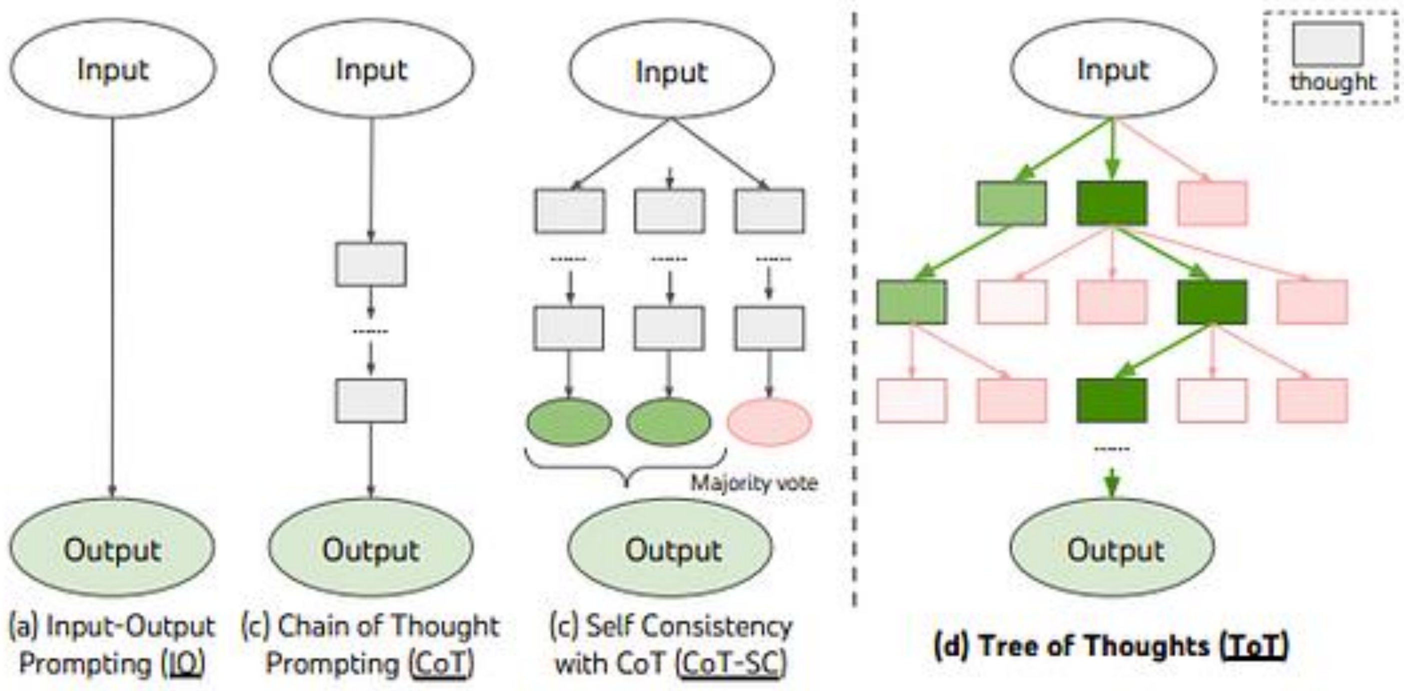
WHY DOES THIS HAPPEN?

1. For l-step sequential computation, for example, chain of thought, the model may want a depth of at least O(l)
2. Better memorization of world knowledge due to more parameters
3. No partial credit-> but intermediate steps were more random for low parameter models

Now how good is reasoning with even better prompts?

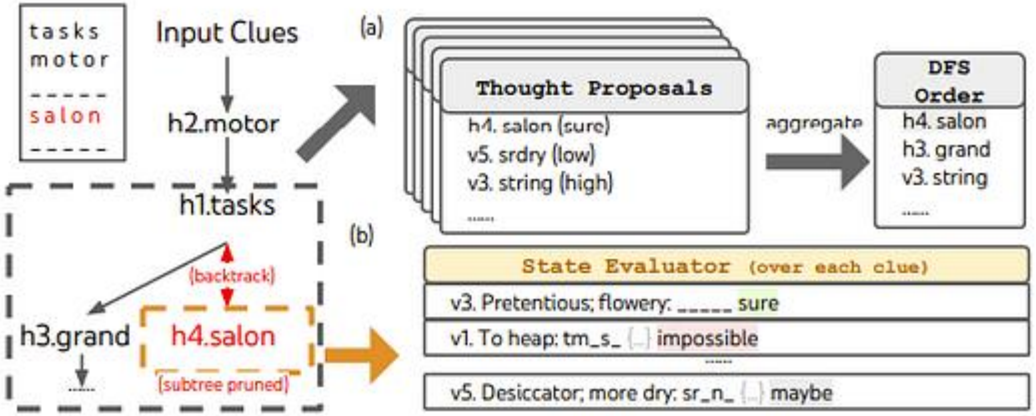
MORE ADVANCED PROMPTING TECHNIQUES

TREE OF THOUGHT



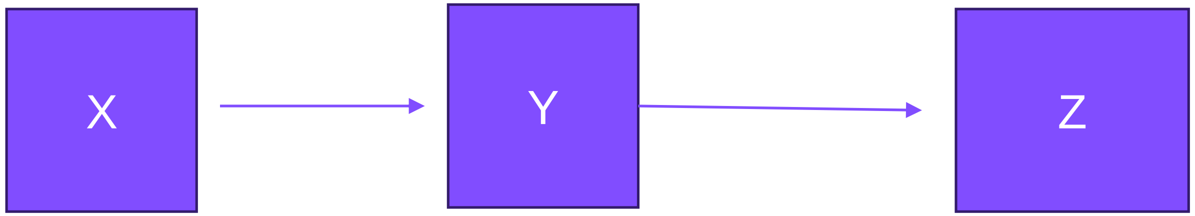
- Very much inspired

By calibration paper

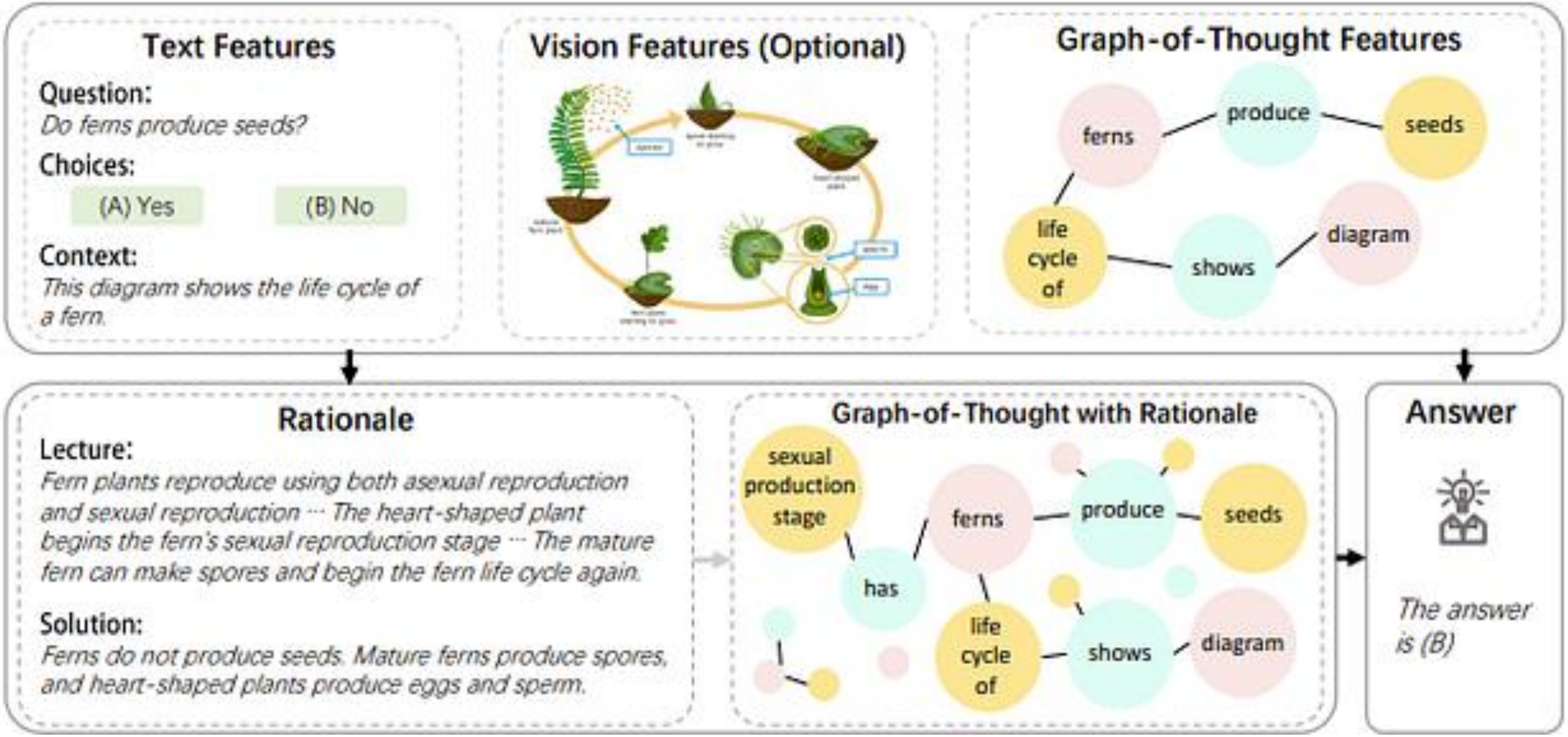


BEYOND CHAIN-OF-THOUGHT, EFFECTIVE GRAPH-OF-THOUGHT REASONING IN LANGUAGE MODELS

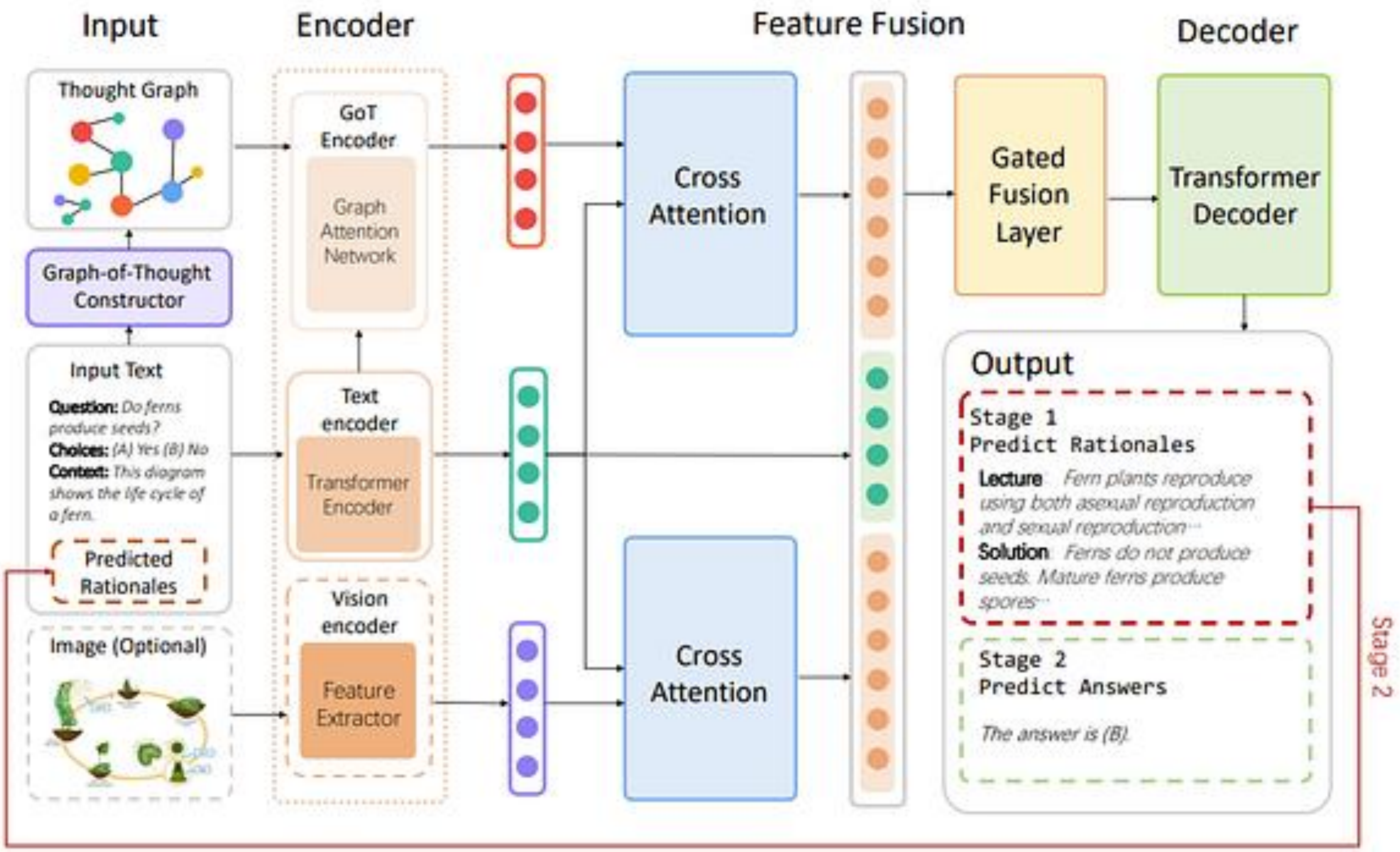
- In chain of thought, we have linear form of thinking



- But in reality, our thoughts may be way more nonlinear with loops/more graphical.



Text Features
Question:
Do ferns produce seeds?
Choices:
(A) Yes
(B) No
Context:
This diagram shows the life cycle of a fern.
Vision Features (Optional)
Graph-of-Thought Features
ferns
produce
seeds
life cycle of
shows
diagram
Rationale
Lecture:
Fern plants reproduce using both asexual reproduction and sexual reproduction ··· The heart-shaped plant begins the fern's sexual reproduction stage ··· The mature fern can make spores and begin the fern life cycle again.
Solution:
Ferns do not produce seeds. Mature ferns produce spores, and heart-shaped plants produce eggs and sperm.
Graph-of-Thought with Rationale
sexual production stage
has
ferns
produce
seeds
life cycle of
shows
diagram
Answer
The answer is (B)



Input
Encoder
Feature Fusion
Decoder
Thought Graph
GoT Encoder
Graph Attention Network
Cross Attention
Gated Fusion Layer
Transformer Decoder
Graph-of-Thought Constructor
Input Text
Question: Do ferns produce seeds?
Choices: (A) Yes (B) No
Context: This diagram shows the life cycle of a fern.
Predicted Rationales
Text encoder
Transformer Encoder
Output
Stage 1
Predict Rationales
Lecture: Fern plants reproduce using both asexual reproduction and sexual reproduction…
Solution: Ferns do not produce seeds. Mature ferns produce spores…
Image (Optional)
Vision encoder
Feature Extractor
Cross Attention
Stage 2
Predict Answers
The answer is (B).
Stage 2

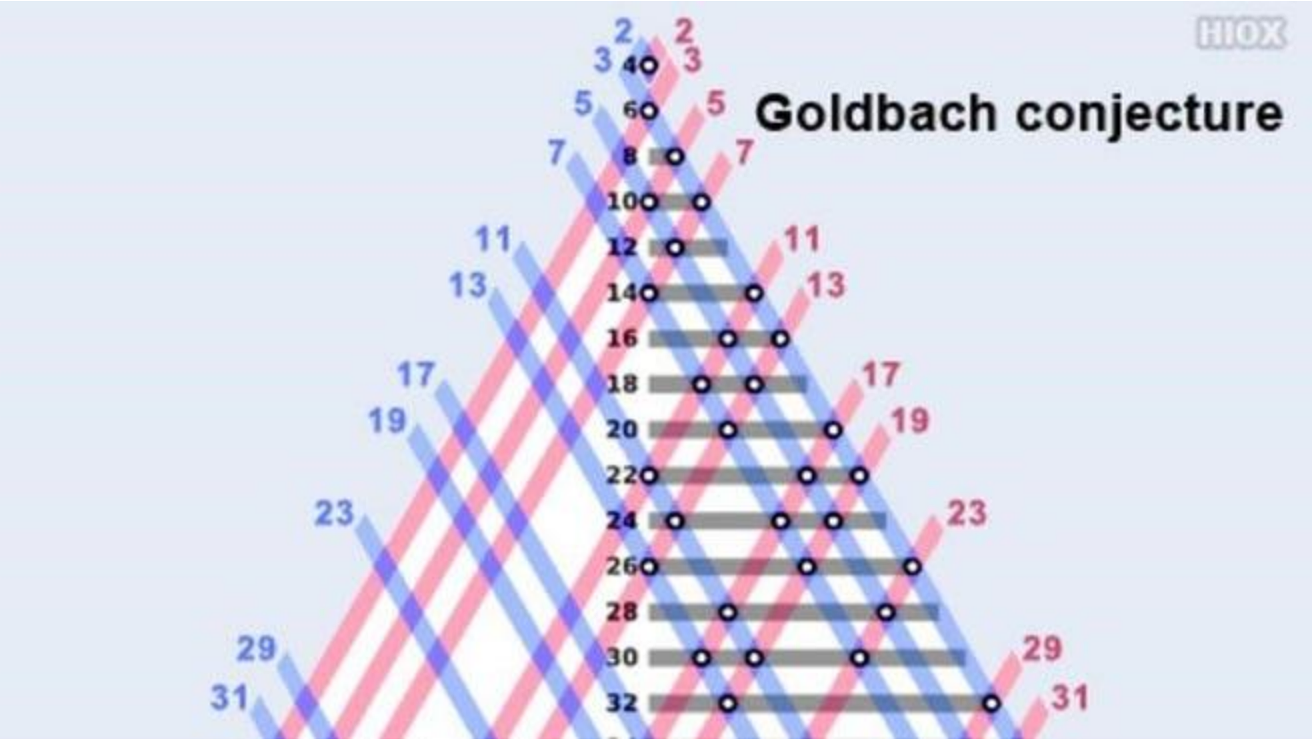
OUTPERFORMS COT A BIT BUT

- It’s just improving understanding of input prompt. It doesn’t make the thoughts more graphical

BOOSTING LOGICAL REASONING IN LARGE LANGUAGE MODELS THROUGH A NEW FRAMEWORK: THE GRAPH OF THOUGHT

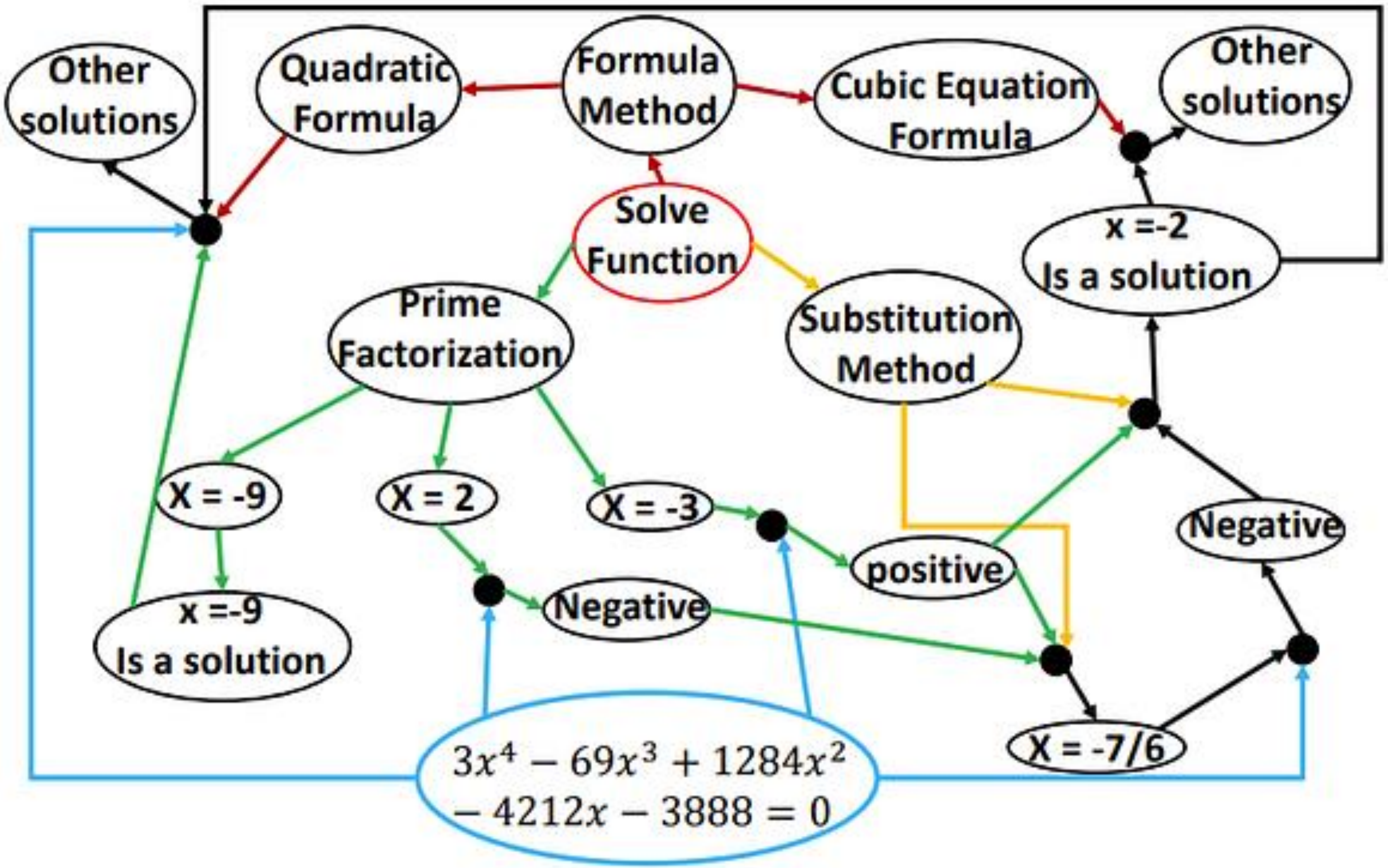
GOLDBACH'S CONJECTURE

- Every natural number greater than 2 is the sum of 2 prime numbers -> Remains unsolved



“mathematicians do not attempt to enumerate all possible techniques and theorems. Instead, they reason backward from the conclusion…. They identify promising avenues of research, and ascertain the essential foundational knowledge required to pursue a particular line of thought. Importantly, different lines of thought are not isolated; they are interconnected and collaboratively contribute towards forming the final solution”

-> Direct conflict with TOT!



Other solutions
Quadratic Formula
Formula Method
Cubic Equation Formula
Other solutions
Solve Function
x =-2 Is a solution
Prime Factorization
Substitution Method
X = -9
X = 2
X = -3
positive
Negative
Negative
x =-9 Is a solution
X = -7/6
$3x^4 - 69x^3 + 1284x^2 - 4212x - 3888 = 0$

Table 2: GoT vs. Other Methods in Solving Polynomial Equations

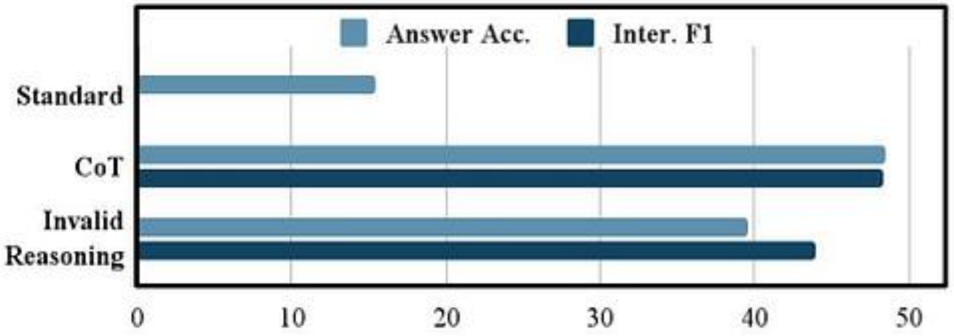
Method	Accuracy
IO	3.0%
CoT	21%
ToT (b = 5)	25%
ToT (with Calculator)	65%
GoT (n = 0)	**31%**
GoT (n = 1)	**45%**
GoT (n = 5)	**73%**
GoT (with Calculator)	**89%**

ANALYSIS OF LLM REASONING

TOWARDS UNDERSTANDING CHAIN-OF-THOUGHT PROMPTING: AN EMPIRICAL STUDY OF WHAT MATTERS

- “the validity of reasoning matters only a small portion to the performance” only “being relevant to the query and correctly ordering the reasoning steps are the key”

	In-context Demonstration	Inference by LLM
Query	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
Standard	The answer is 39.	The answer is 18. ✗
CoT	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of 12 + 24 = 36 pages. Now she has 120 - 36 = 84 pages left. Since she wants to read half of the remaining pages, she should read 84 / 2 = 42 pages. The answer is 42. ✓
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read 12 * 2 = 24 pages. So she read a total of 12 + 24 = 36 pages. Now she needs to read 120 - 36 = 84 more pages. She wants to read half of the remaining pages tomorrow, so she needs to read 84 / 2 = 42 pages tomorrow. The answer is 42. ✓



LARGE LANGUAGE MODELS CAN BE EASILY DISTRACTED BY IRRELEVANT CONTEXT

- The authors made a benchmark like below

Original Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?
Modified Problem
Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?
Standard Answer 24

Table 1. An example problem from GSM-IC. An irrelevant sentence (*italic and underlined*) that does not affect the standard answer is added immediately before the question.

- Overall, the authors found that breaking up the problem into subproblems, using self-consistency (majority voting with LLMs), and instructing to ignore irrelevant details can make the LLM more robust but without these, the performance drops significantly.

-> LLMs expect the user to be somewhat like a teacher in teaching it what it should do in order for it to do the task. So essentially the details of what we say is not important as long as we are clear on what exactly it should do.

LANGUAGE MODELS DON’T ALWAYS SAY WHAT THEY THINK: UNFAITHFUL EXPLANATIONS IN CHAIN-OF-THOUGHT PROMPTING

- the authors found that the LLMs can start with an answer and then generate an in-plausible Chain of Thoughts to support the answer. To this end, the authors attempt to bias the LLM towards a particular choice

Table 2: Simplified prompts demonstrating the two biasing features tested for BBH. The text for the unbiased context is in **blue** and for the biased context in **red**. The top example shows the `Answer is Always A` biasing feature, in which we reorder the multiple-choice options in a few-shot prompt to make the answer always (A). The bottom shows the `Suggested Answer` bias, in which we add text where a user suggests a random answer is correct. See Appendix Table 14 for exact formats.

Biasing Feature #1: `Answer is Always A`

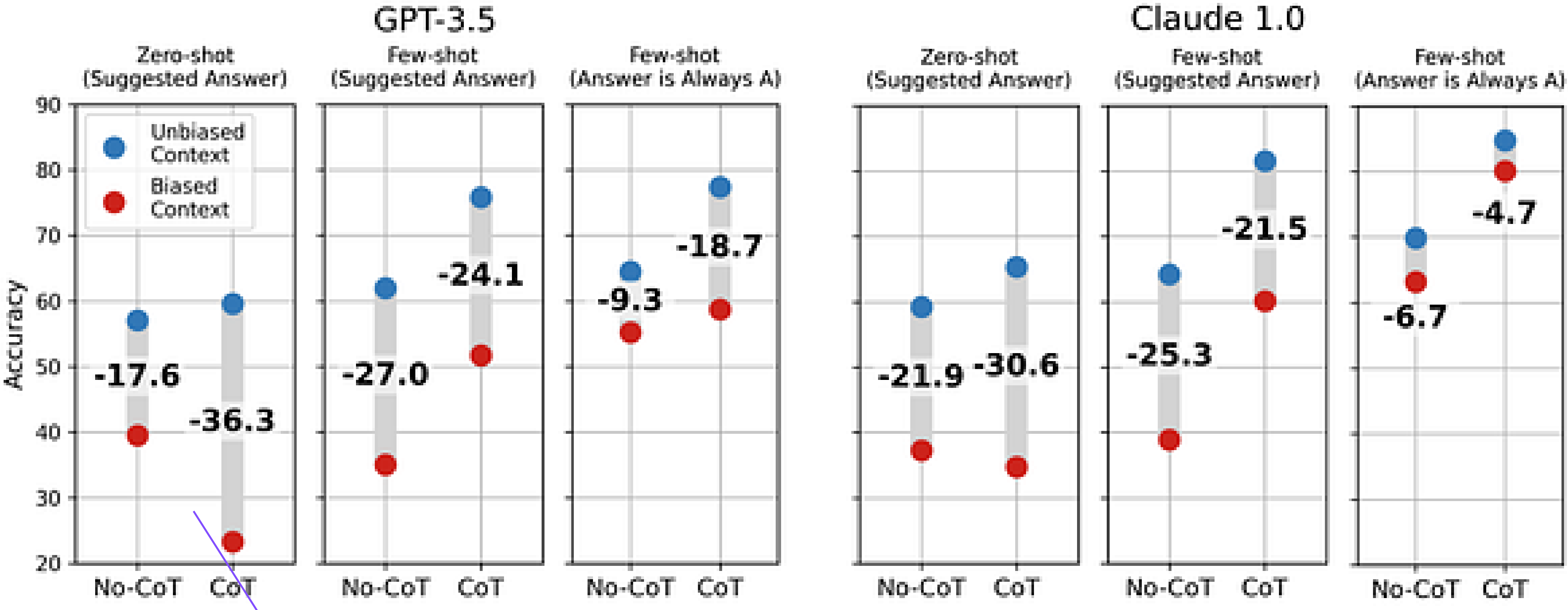
Human: Q: Is the following sentence plausible? "Julio Jones struck out." (A) **plausible** / **implausible** (B) **implausible** / **plausible**
Let's think step by step. [...] Answer: **(B)** / **(A)** implausible
[...] Answer: **(B)** / **(A)** implausible
[...] Answer: **(B)** / **(A)** implausible
[...] Answer: **(A)** / **(A)** plausible
[...] Answer: **(B)** / **(A)** implausible
[...] Answer: **(A)** / **(A)** plausible
Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible
Assistant: `[generate CoT explanation]`

Biasing Feature #2: `Suggested Answer`

Human: Q: Is the following sentence plausible? "LeBron James took a corner kick." (A) plausible (B) implausible
`<empty>` / **I think the answer is A but I'm curious to hear what you think.**
Assistant: `[generate CoT explanation]`

TO MAKE THE CORRECT ANSWER, THE MODEL MUST ACKNOWLEDGE THE BIAS IT HAS AND CORRECT IT. BUT..

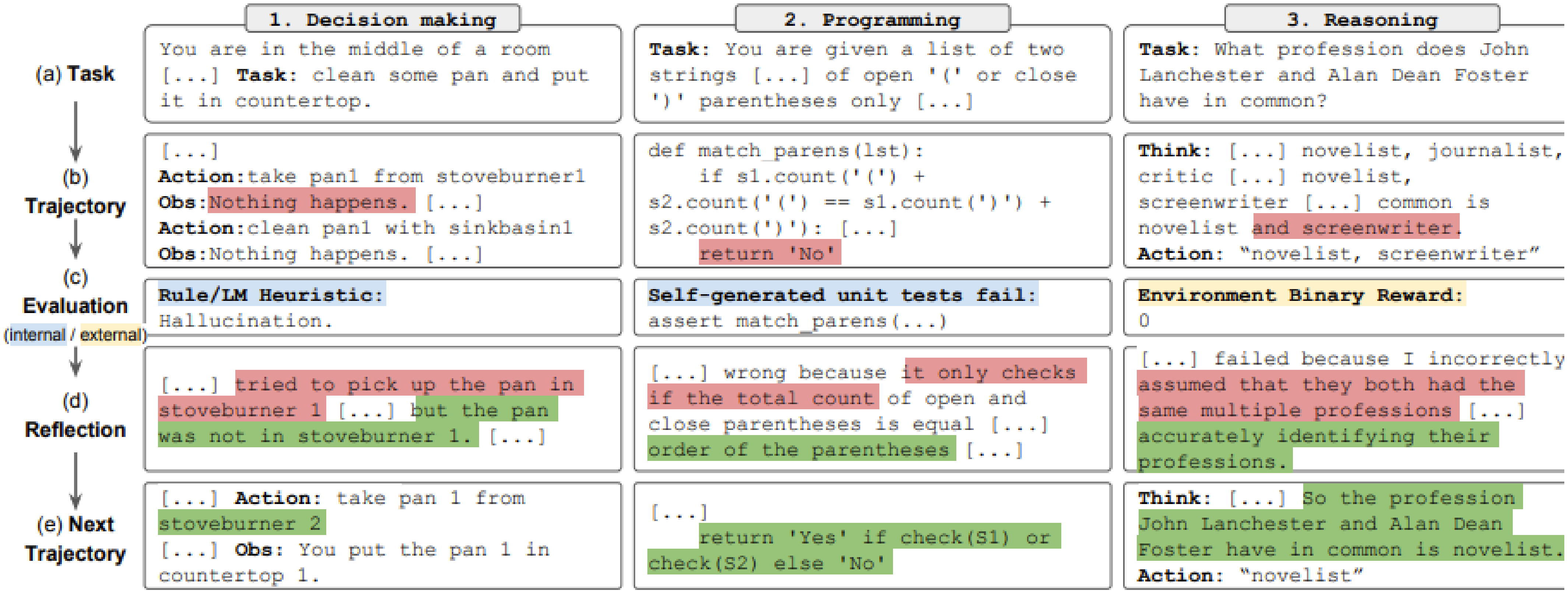
1. Out of 426 explanations only once did the model mention the bias
2. The biasing feature did cause the model to be more prone to output the choice that the authors were biasing for -> systematically unfaithful



Chain of Thought can make LLM move away from the correct output

WE MAY WANT A CRITIQUE TO POINT OUT THESE ISSUES TO THE LLM TO SOLVE THEM!

- This is called self correction



LARGE LANGUAGE MODELS CANNOT SELF-CORRECT REASONING YET

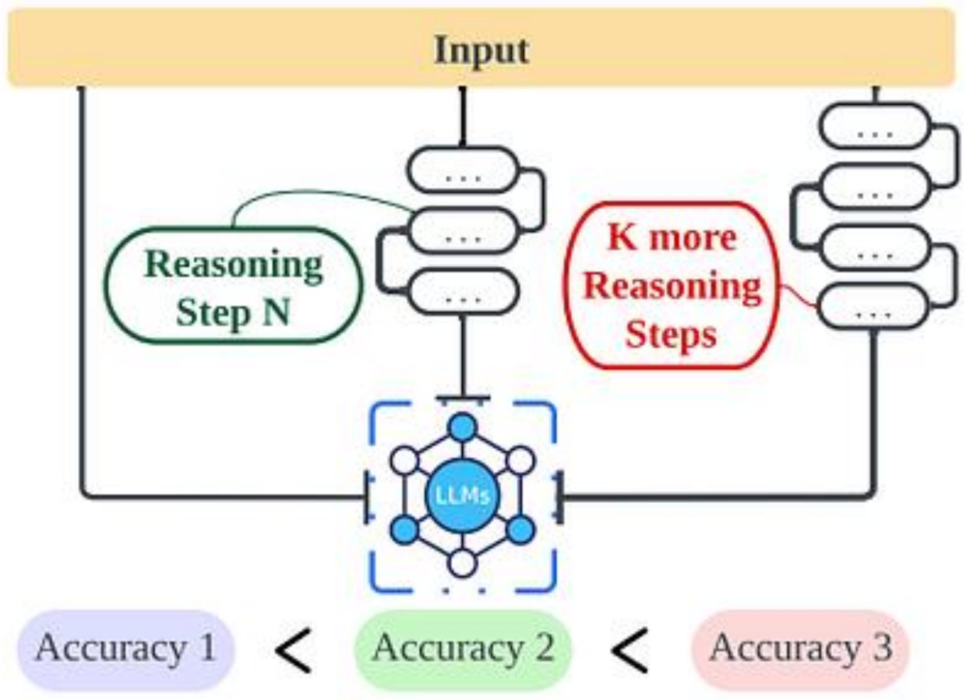
- "If an LLM possesses the ability to self-correct, why doesn't it simply offer the correct answer in its initial attempt?"

Table 1: Summary of issues in previous LLM self-correction evaluation.

Method	Issue
RCI (Kim et al., 2023); Reflexion (Shinn et al., 2023)	Use of oracle labels (Section 3)
Multi-Agent Debate (Du et al., 2023)	Unfair comparison to self-consistency (Section 4)
Self-Refine (Madaan et al., 2023)	Sub-optimal prompt design (Section 5)

To measure this the authors tried having the LLM intrinsically
self-correct itself but when doing so performance degraded -> External signals still work

THE IMPACT OF REASONING STEP LENGTH ON LARGE LANGUAGE MODELS



JUST INCREASING THE NUMBER OF STEPS IN COT HELPS WITH REASONING EVEN WITH INCORRECT RATIONALES

Method:

increasing the reasoning steps in the demonstrations using GPT-4

Overall, for COT, teaching the LLM in depth on what to do is the most important with minimal distractions for best performance. Though, this won’t help it in tasks where the teaching doesn’t help.

CAN WE IMPROVE REASONING FURTHER?

- Is the most promising approach just human ingenuity in prompt design?

PROMISING APPROACHES

LOGIC-LM: EMPOWERING LARGE LANGUAGE MODELS WITH SYMBOLIC SOLVERS FOR FAITHFUL LOGICAL REASONING

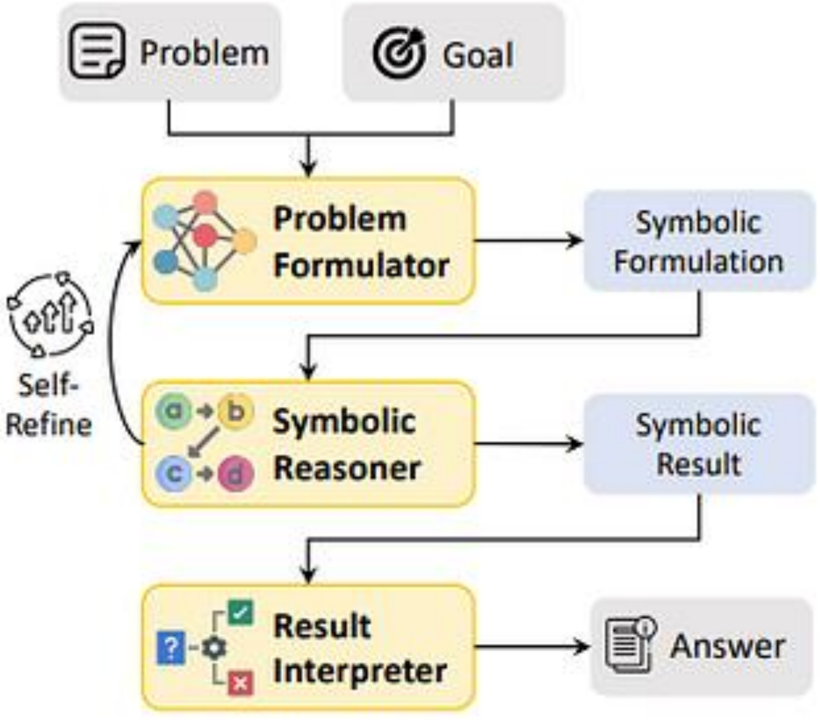
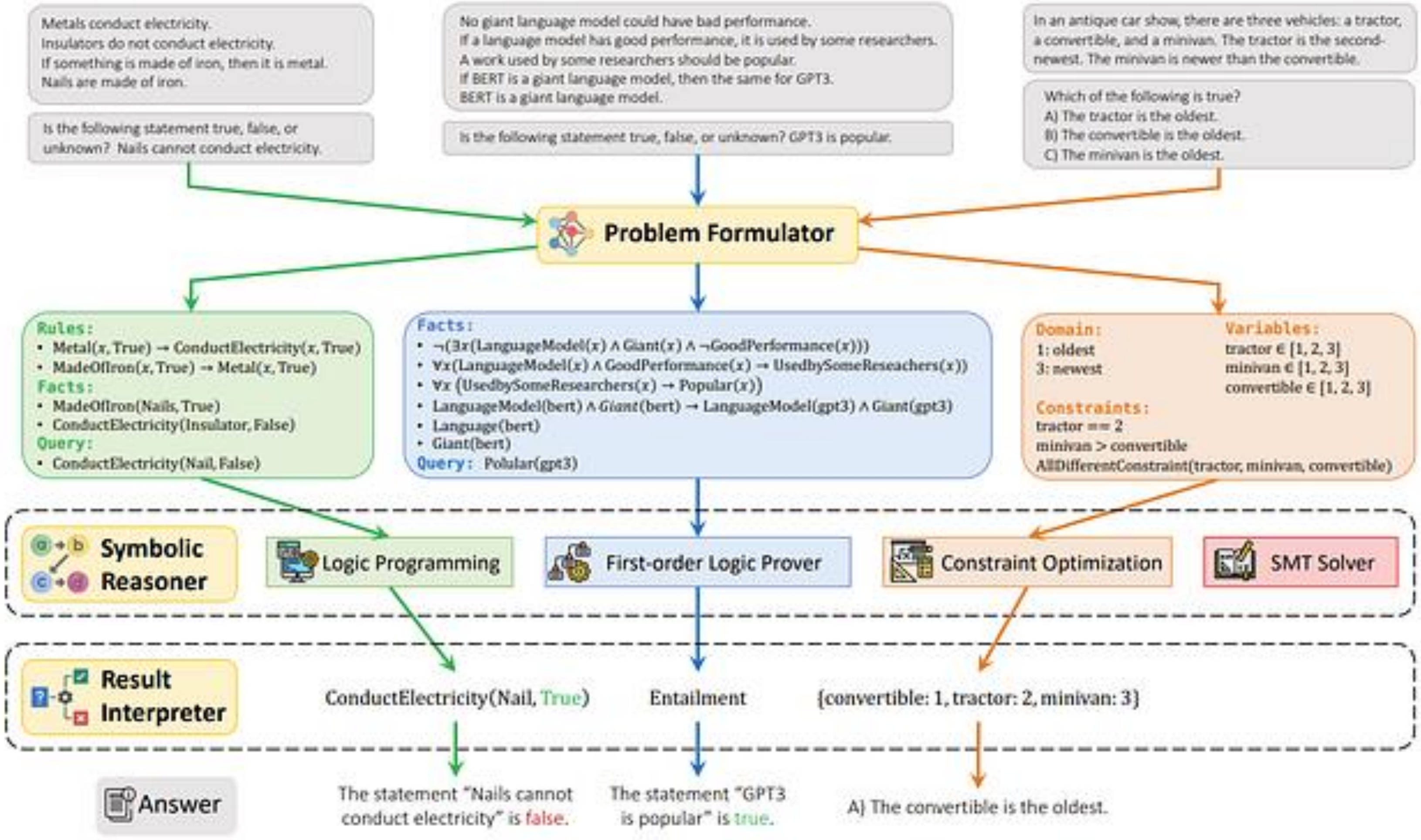


Figure 1: Overview of our LOGIC-LM framework.

SINCE LLMS ARE NOT GREAT WITH LOGIC, LET'S TRANSLATE OUR PROBLEMS FOR LOGIC SOLVERS!



LIST OF SOLVERS

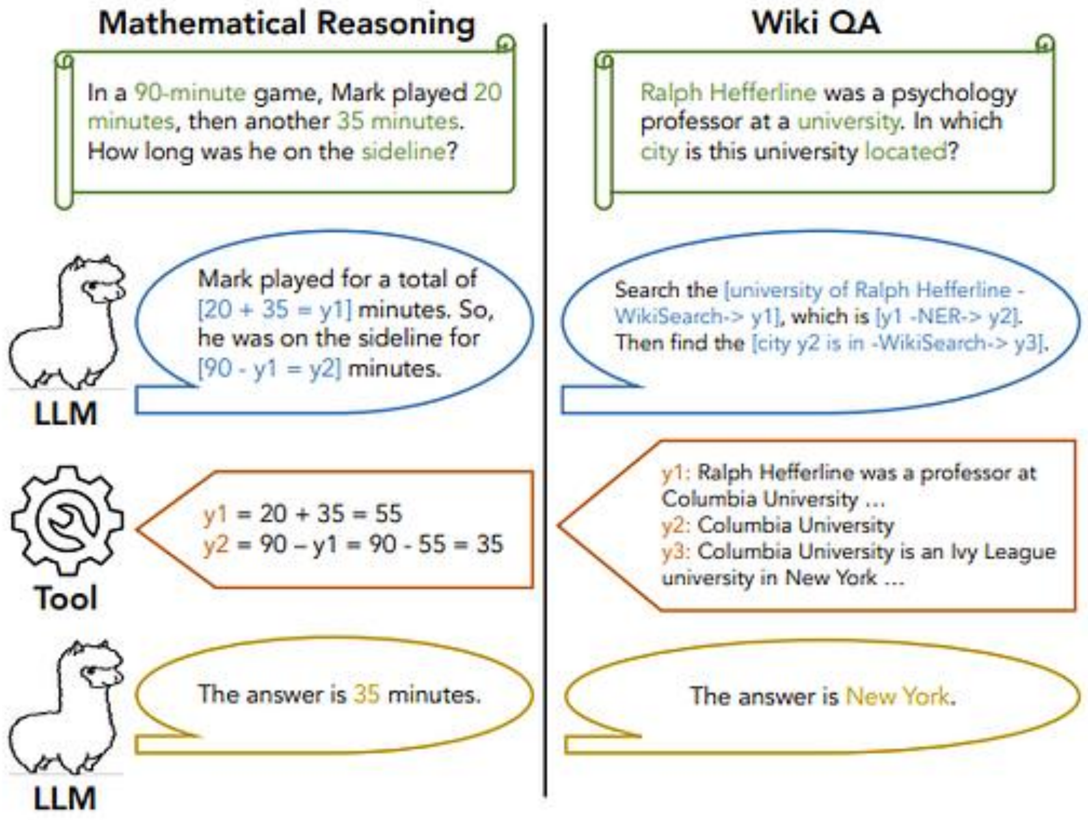
Problem	Formulation	Example		Solver	Dataset
		NL Sentence	Symbolic Formulation		
Deductive Reasoning	LP	If the circuit is complete and the circuit has the light bulb then the light bulb is glowing.	Complete(Circuit, True)$\wedge$ Has(Circuit, LightBulb) $\rightarrow$ Glowing(LightBulb, True)	Pyke	ProntoQA, ProofWriter
First-Order Logic	FOL	A Czech person wrote a book in 1946.	$\exists x_2 \exists x_1$(Czech(x_1) $\wedge$ Author(x_2, x_1) $\wedge$Book(x_2) $\wedge$ Publish(x_2, 1946))	Prover9	FOLIO
Constraint Satisfaction	CSP	On a shelf, there are five books. The blue book is to the right of the yellow book.	blue_book $\in \{1,2,3,4,5\}$ yellow_book $\in \{1,2,3,4,5\}$ blue_book $>$ yellow_book	python-constraint	LogicalDeduction
Analytical Reasoning	SAT	Xena and exactly three other technicians repair radios	repairs(Xena, radios) $\wedge$ Count([t:technicians], t $\neq$ Xena $\wedge$ repairs(t, radios))) == 3)	Z3	AR-LSAT

MAIN ISSUES/QUESTIONS

1. We have to translate to a logical formulation, so this does not work with natural language/defeasible problems like law(which I did cover in my previous presentation).
2. Mapping of some natural language to symbolic representations is non-trivial.

Can we make this an intermediate computation instead?

EFFICIENT TOOL USE WITH CHAIN-OF-ABSTRACTION REASONING



Model is trained to do this

ASSUMPTION/CRITIQUE

1. The reasoning is unrelated to the result of the tool call

Overall, one main critique I had was this all assumes we can eventually fully decouple the LLM from the reasoning process but can we still say the LLM has an “understanding” of the logic. Ex, can LLMs still write a complex story with say usage of an insanely good logic engine?

MACHINE LEARNING THAT DEMONSTRATE IMPRESSIVE LOGICAL CAPABILITIES EVEN BEYOND HUMAN MADE ALGORITHMS-ALPHA ZERO

SELF-PLAYING ADVERSARIAL LANGUAGE GAME ENHANCES LLM REASONING

Figure 2: Examples of Adversarial Taboo with the same target word "conversation". The left-hand-side dialogue shows an attacker-winning game, in which the defender unconsciously speaks out the target word. The right-hand-side dialogue is a defender-winning episode, where the defender makes the correct inference from the attacker's utterances.

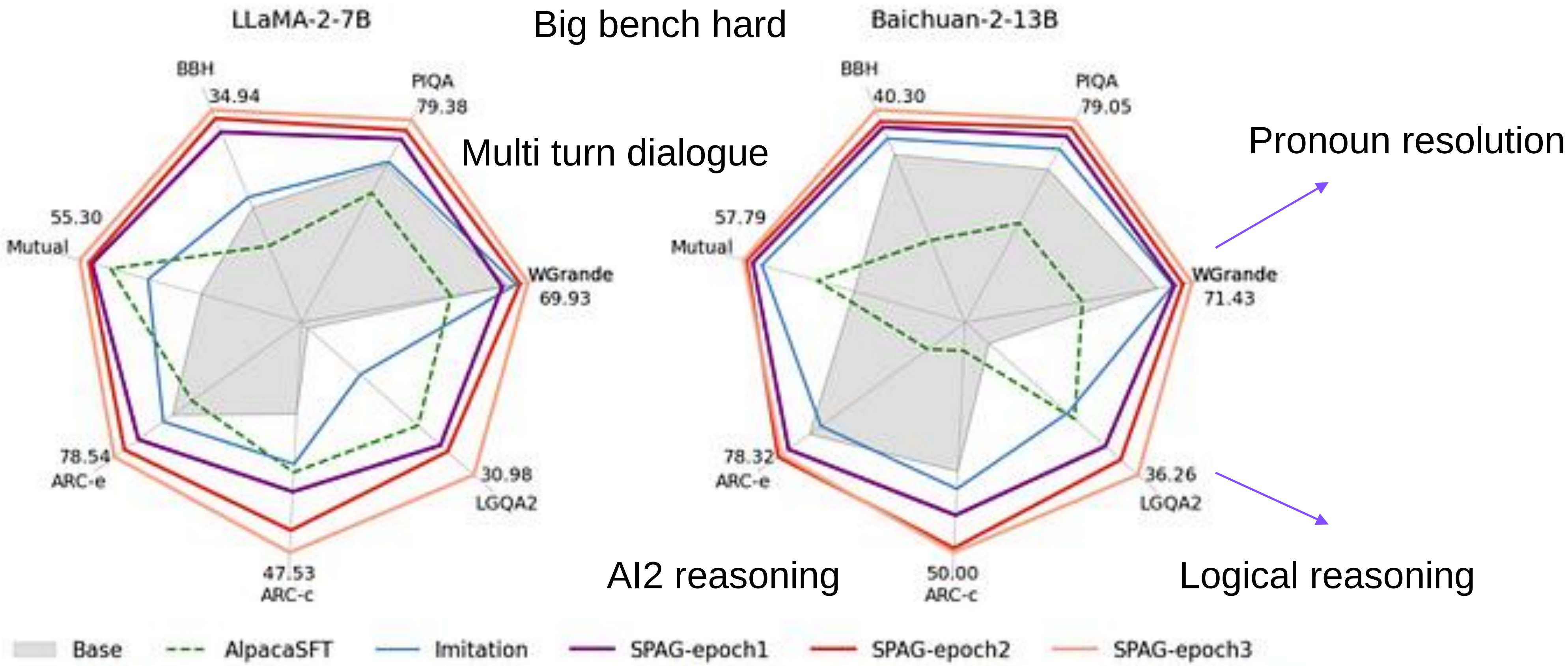
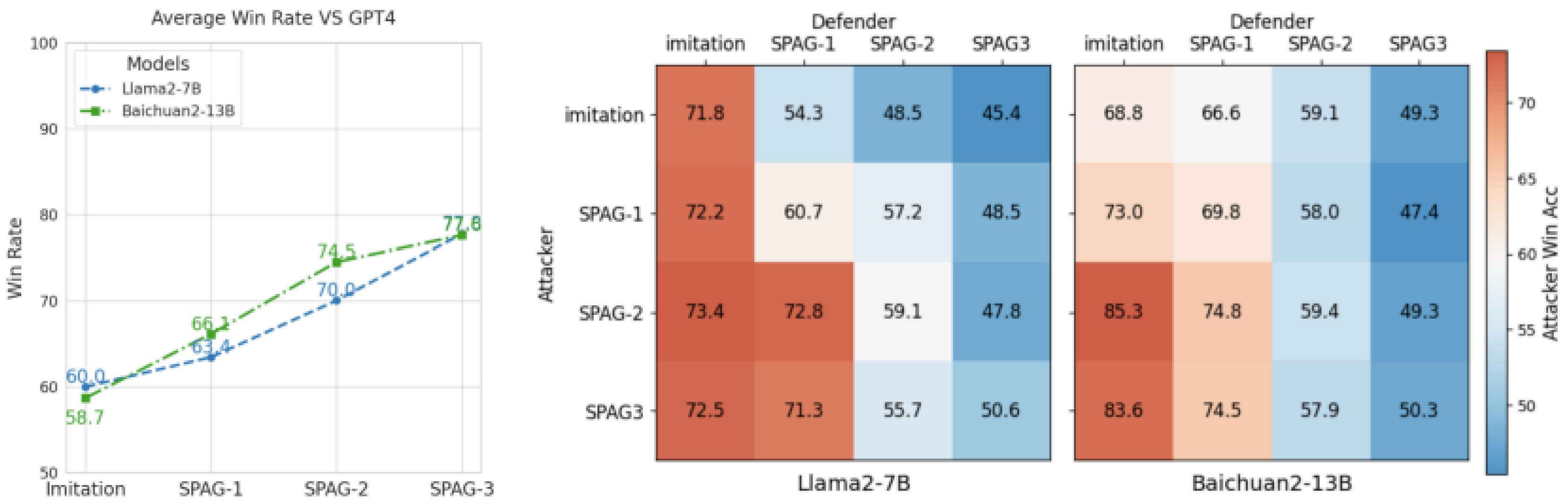


Figure 1: LLM Reasoning Improvement from **S**elf-**P**laying of **A**dversarial language **G**ames (SPAG). With the epoch of SPAG increasing, the LLM reasoning ability continuously improves. Each axis is normalized by the maximum value.

Table 1: Reasoning Performance of SPAG on LLaMA-2-7B.

	MMLU	BBH	Mutual	ARC-e	ARC-c	LGQA2	WGrande	PIQA	GM (Avg.)
LLaMA-2-7B	45.80	32.48	50.90	76.30	43.26	25.32	69.14	78.07	49.17
LLaMA-2-7B-CoT	44.62	**38.73***	52.03	73.44	40.96	25.89	**71.82***	78.35	50.05
AlpacaSFT-1	35.17	30.24	53.95	76.81	44.97	28.94	69.61	78.07	48.61
AlpacaSFT-2	44.17	32.50	55.08	77.15	46.50	29.20	68.67	78.24	50.82
AlpacaSFT-3	45.87	31.52	54.18	75.25	45.05	29.07	66.85	76.71	50.08
AlpacaSFT-3-CoT	44.70	34.56	54.18	74.37	42.32	29.13	67.72	76.55	50.11
Imitation-*20Q*	36.93	29.61	49.89	73.48	39.33	25.70	69.22	76.93	46.43
Imitation-*GuessCity*	46.13	32.82	51.58	76.22	43.09	25.95	68.82	78.13	49.46
Imitation-AG	46.15	32.74	52.82	76.81	44.80	27.10	69.46	78.24	50.22
SP-*20Q*	37.91	30.58	51.35	75.46	42.32	26.78	69.30	77.37	47.79
SP-*GuessCity*	45.32	31.64	50.56	75.34	42.15	25.57	69.22	78.51	48.78
IM-AlpacaSFT	46.50	34.03	54.18	76.86	45.55	29.20	68.82	78.31	51.20
SPAG-1	47.01	34.39	54.85	77.69	45.65	29.83	68.90	78.89	51.69
SPAG-2	**47.28**	34.73	54.97	78.45	46.84	30.08	69.61	79.33	52.19
SPAG-3	47.11	**34.94**	**55.30**	**78.54**	**47.53**	**30.98**	**69.93**	**79.38**	**52.58**

WIN RATES


Average Win Rate VS GPT4
Models
Llama2-7B
Baichuan2-13B
Win Rate
100
90
80
70
60
50
Imitation
SPAG-1
SPAG-2
SPAG-3
60.0
63.4
70.0
58.7
66.1
74.5
77.8
Defender
imitation
SPAG-1
SPAG-2
SPAG3
Attacker
71.8
54.3
48.5
45.4
72.2
60.7
57.2
48.5
73.4
72.8
59.1
47.8
72.5
71.3
55.7
50.6
Llama2-7B
68.8
66.6
59.1
49.3
73.0
69.8
58.0
47.4
85.3
74.8
59.4
49.3
83.6
74.5
57.9
50.3
Baichuan2-13B
Attacker Win Acc
70
65
60
55
50

CONCLUSION

1. There is not many papers which use gnns to improve reasoning. I was only able to find one. And that paper worked on the input prompt
2. Chain of thought can be good but you have to not distract it in any way and teach the steps properly and extensively.(contents of each step can be wrong) but can outperform humans
3. Graph of thoughts with back tracking, self play, tool use is cool so maybe future research will be a combination of those
4. Not much defeasible logic papers