ACL 2024

Boosting Language Models Reasoning with Chain-of-Knowledge Prompting

Jianing Wang[1*], Qiushi Sun[2*] , Xiang Li[1†], Ming Gao[1]

[1] East China Normal University [2]The University of Hong Kong

[*] Equal Contribution [†] Corresponding Author

Background & Motivations

Reasoning with LLMs

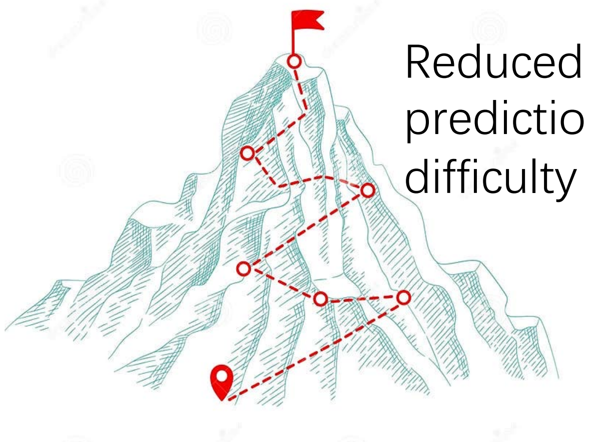
- The prosperity of LLMs
  - GPT
  - Gemini
  - Claude
  - LLaMA Series
  - MOSS/InternLM/Qwen
  - …
- Strategies
  - In-Context Learning
  - Task Decomposition

Reasoning with LLMs

The success of LLMs on reasoning: spontaneously decompose the complex problem into intermediate reasoning chains

"Let's think step by step"

Direct target: $P(y|x)$

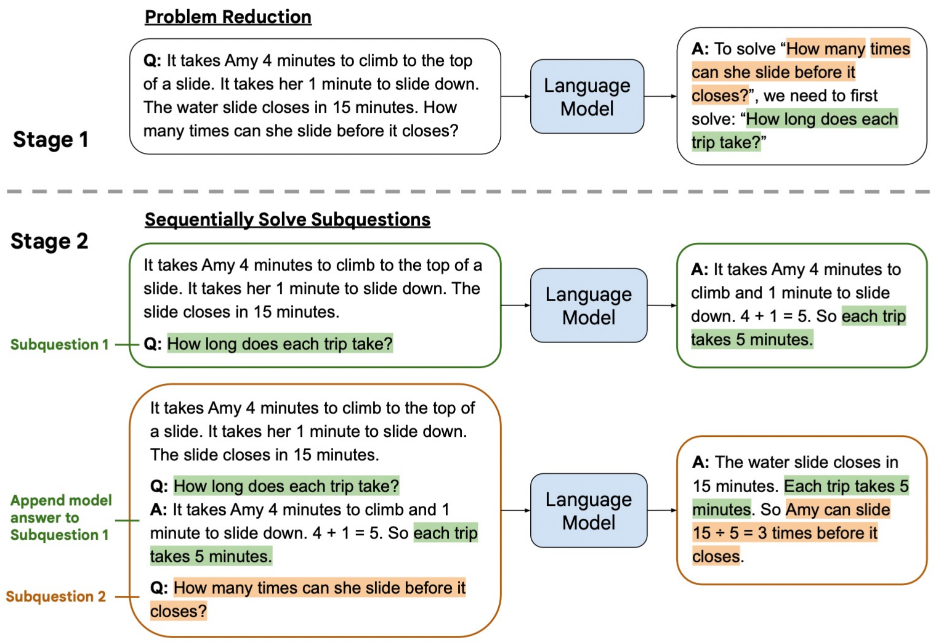


Reduced prediction difficulty

Indirect target: $P(y|x) = P(y|z_n) \cdots P(z_1|x)$

Reasoning with LLMs

The success of LLMs on reasoning: spontaneously decompose the complex problem into intermediate reasoning chains



Q: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

A: To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down. 4 + 1 = 5. So each trip takes 5 minutes.

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

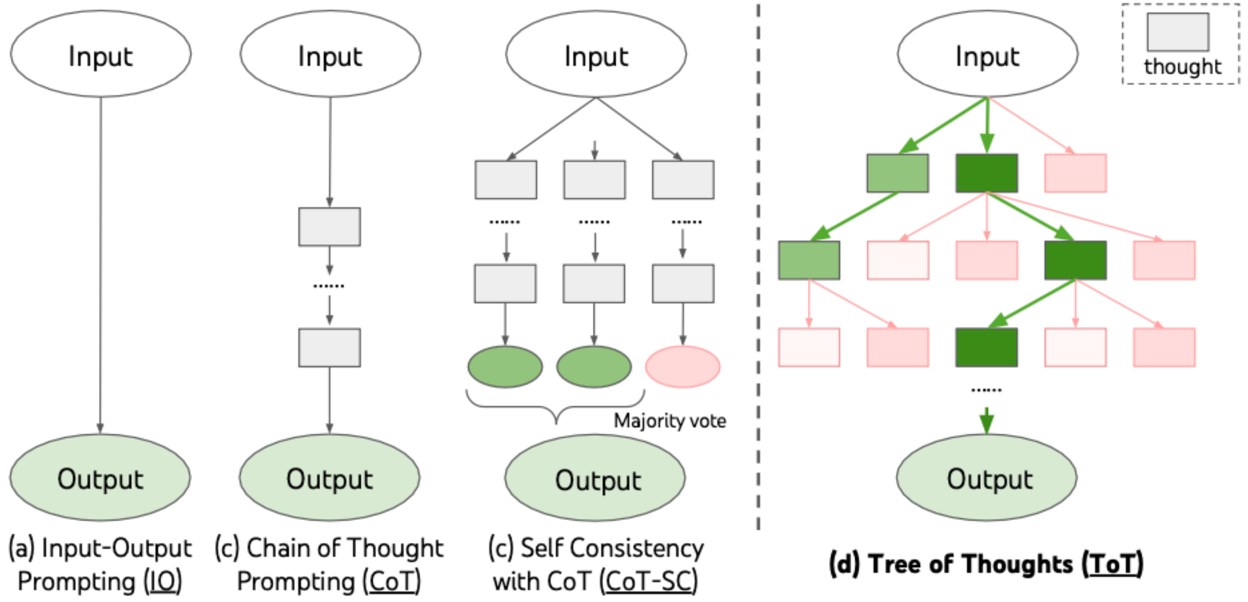
Q: How long does each trip take?

A: It takes Amy 4 minutes to climb and 1 minute to slide down. 4 + 1 = 5. So each trip takes 5 minutes.

Q: How many times can she slide before it closes?

A: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide 15 ÷ 5 = 3 times before it closes.

Least-to-Most Decomposition



(a) Input-Output Prompting (IO)

(c) Chain of Thought Prompting (CoT)

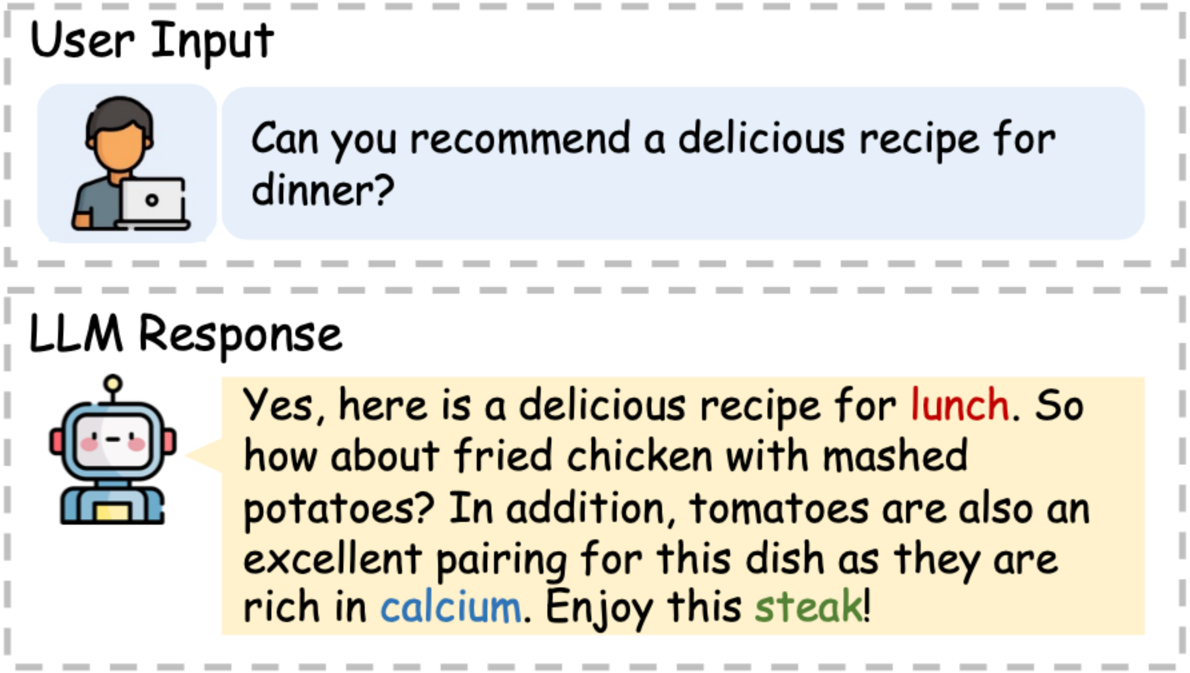
(c) Self Consistency with CoT (CoT-SC)

(d) Tree of Thoughts (ToT)

Chain/Tree-of-Thought

Problems

- ■ Hallucinated Output
  - □ Input-Conflicting: Unfaithfulness of Input-Output
  - □ Fact-Conflicting: Unfactual of Output-Facts



User Input

Can you recommend a delicious recipe for dinner?

LLM Response

Yes, here is a delicious recipe for lunch. So how about fried chicken with mashed potatoes? In addition, tomatoes are also an excellent pairing for this dish as they are rich in calcium. Enjoy this steak!

Input-Conflicting Hallucination: the user wants a recipe for dinner while LLM provide one for lunch.

Fact-Conflicting Hallucination: tomatoes are not rich in calcium in fact.

Hallucination Problem in LLMs' Reasoning:
Cannot verify the reliability of rationales!

Tackling this issue

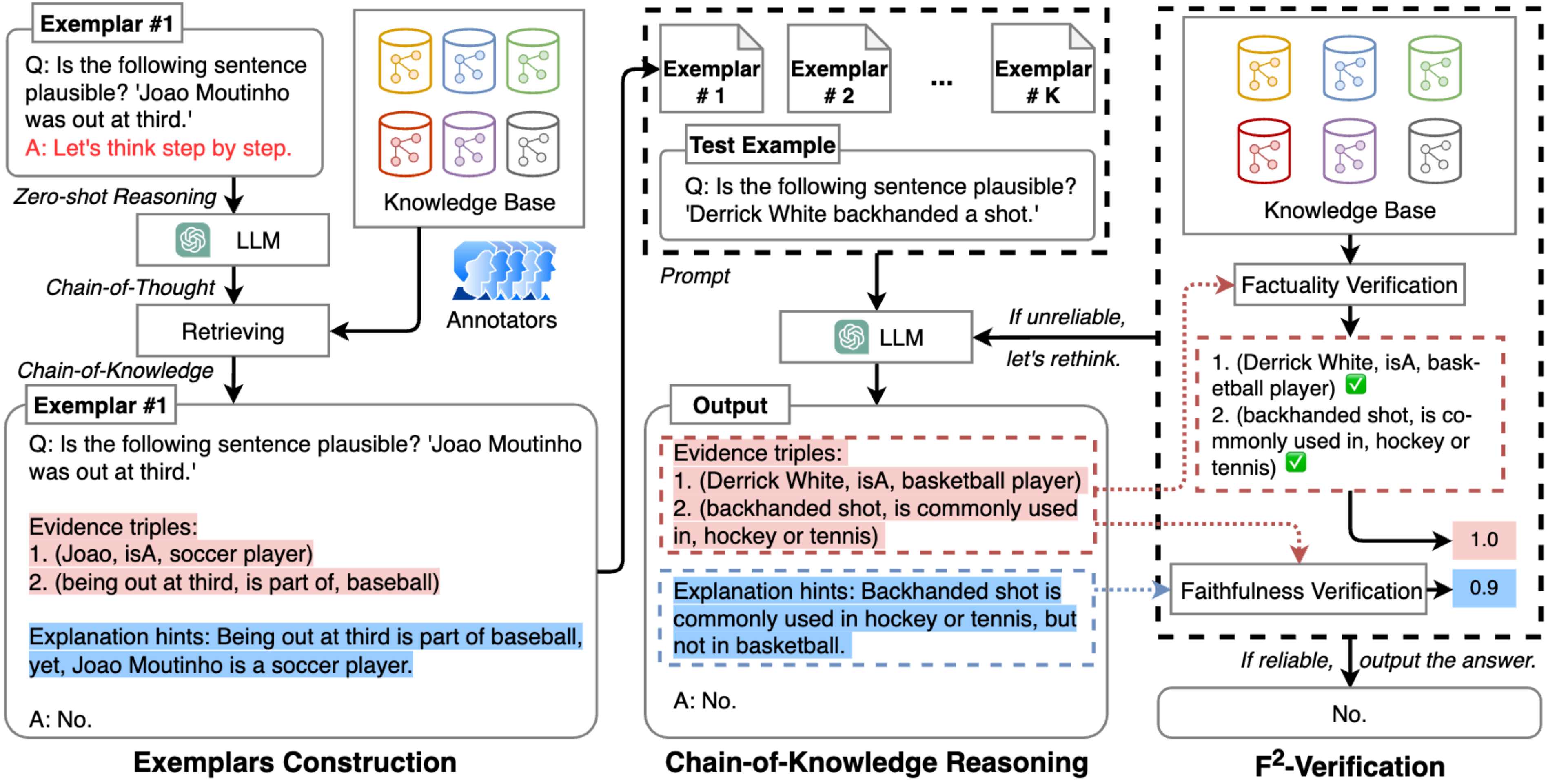
- Recap: How LLMs are built:
  - □ Pre-training on ultra-large-scale corpus: Learn about prior information.
  - □ Supervised fine-tuning on instruction-like data: Learn about instruction-following capabilities.
  - □ Aligning with preferences via RLHF, DPO, etc.: Learn about reliable response.
  - □ Reasoning on complex problems: Learn about how to solve a task.

Have We Fully Utilized the Knowledge Learned by the Model?

Elicit It to Boost LLM Reasoning!

Methodology

Chain-of-Knowledge (CoK) Prompting



Exemplar #1
Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'
A: Let's think step by step.
Zero-shot Reasoning
LLM
Chain-of-Thought
Retrieving
Knowledge Base
Annotators
Chain-of-Knowledge

Exemplar #1

Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'

Evidence triples:
1. (Joao, isA, soccer player)
2. (being out at third, is part of, baseball)

Explanation hints: Being out at third is part of baseball, yet, Joao Moutinho is a soccer player.

A: No.

Q: Is the following sentence plausible? 'Derrick White backhanded a shot.'

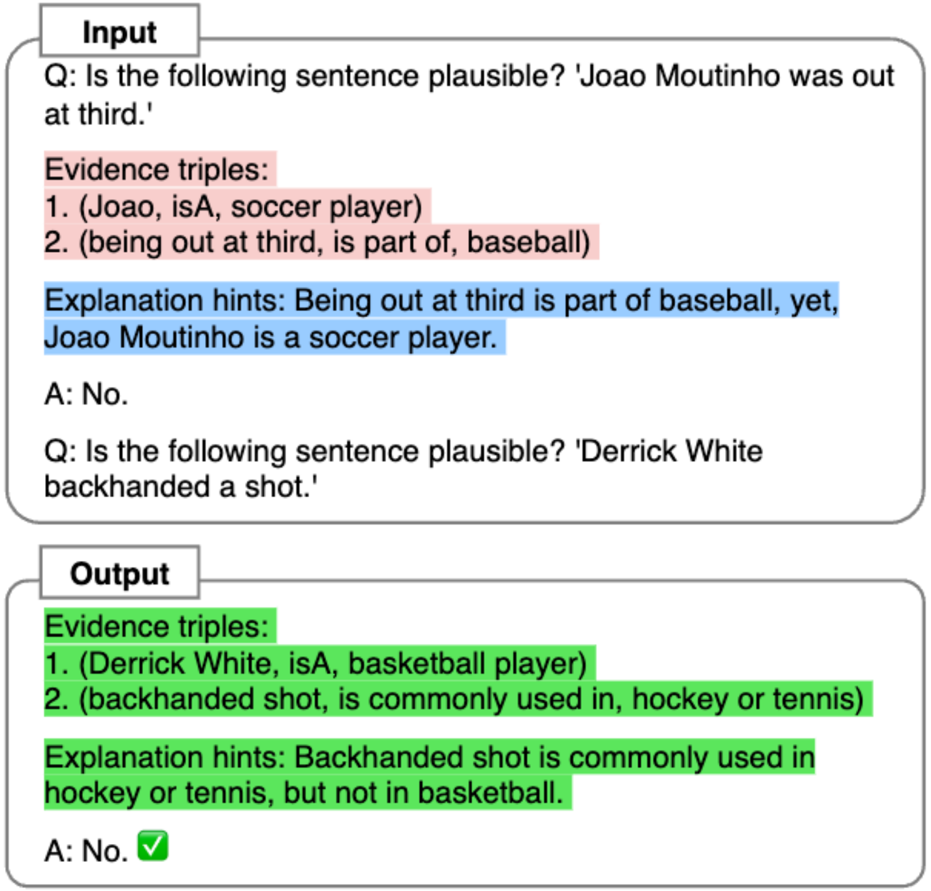
Evidence triples:
1. (Derrick White, isA, basketball player)
2. (backhanded shot, is commonly used in, hockey or tennis)

Explanation hints: Backhanded shot is commonly used in hockey or tennis, but not in basketball.

A: No.

1. (Derrick White, isA, basketball player) ✅
2. (backhanded shot, is commonly used in, hockey or tennis) ✅

Triples and Hints



Input

Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'

Evidence triples:
1. (Joao, isA, soccer player)
2. (being out at third, is part of, baseball)

Explanation hints: Being out at third is part of baseball, yet, Joao Moutinho is a soccer player.

A: No.

Q: Is the following sentence plausible? 'Derrick White backhanded a shot.'

Output

Evidence triples:
1. (Derrick White, isA, basketball player)
2. (backhanded shot, is commonly used in, hockey or tennis)

Explanation hints: Backhanded shot is commonly used in hockey or tennis, but not in basketball.

A: No. ✅

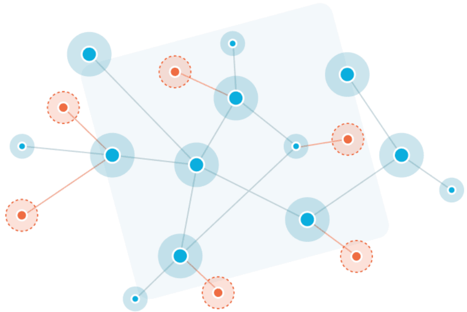
(c) Ours: Chain-of-Knowledge Prompting

- **Evidence Triples**----Structure rationale

(**h**, **r**, **t**): **h** denotes head entity, **r** denotes relation, **t** denotes tail entity.

- **Explanation Hints**----Textual rationale

LLMs can be elicited to generate both structured and textual rationales.



Faithfulness and Factuality

Question: Can the knowledge generated by the model be truly reliable?

Perhaps not, we need to refine and process this knowledge further.

Faithfulness and Factuality

F²-Verification: Faithfulness Score

Leverage SimCSE to calculate the similarity between **Evidence Triples** and **Explanation Hints**.

$$s_u(\widetilde{T}, \widetilde{H}|\widetilde{X}) = \mathrm{SimCSE}(||_{j=1}(T_j), \widetilde{H}),$$

$\widetilde{T} = \{(s_j, r_j, o_j)\}_{j=1}^{L}$: Evidence Triples

$\widetilde{H}$: Explanation Hints

$\widetilde{X}$: Test Input

Faithfulness and Factuality

F²-Verification: Factuality Score

Leverage external knowledge graph to calculate the correctness of each evidence triple.

- Exactly Matching: If the generated triple can be found in KG, we can assign score 1.0;
- Implicit Matching: If not found in KG, we can calculate the energy score (smaller than 1.0)

$$d(s_j, r_j, o_j|\mathcal{G}) = ||\mathbf{s}_j^{(r,c)} + \mathbf{r}^c - \mathbf{o}_j^{(r,c)}||_2^2 + \alpha||\mathbf{r}^c - \mathbf{r}_j||_2^2$$

Faithfulness and Factuality

F²-Verification: Factuality Score

Leverage external knowledge graph to calculate the correctness of each evidence triple.

- Exactly Matching: If the generated triple can be found in KG, we can assign score 1.0;
- Implicit Matching: If not found in KG, we can calculate the energy score (smaller than 1.0)
- Merge them:

$$s_v(\widetilde{T}|\widetilde{X},\mathcal{G}) = \frac{1}{|\widetilde{T}|}\sum_{(s_j,r_j,o_j)\in\widetilde{T}} \Big[\mathbb{I}\big((s_j,r_j,o_j)\in\mathcal{G}\big) + \big(1-\mathbb{I}((s_j,r_j,o_j)\in\mathcal{G})\big)d(s_j,r_j,o_j|\mathcal{G})\Big]$$

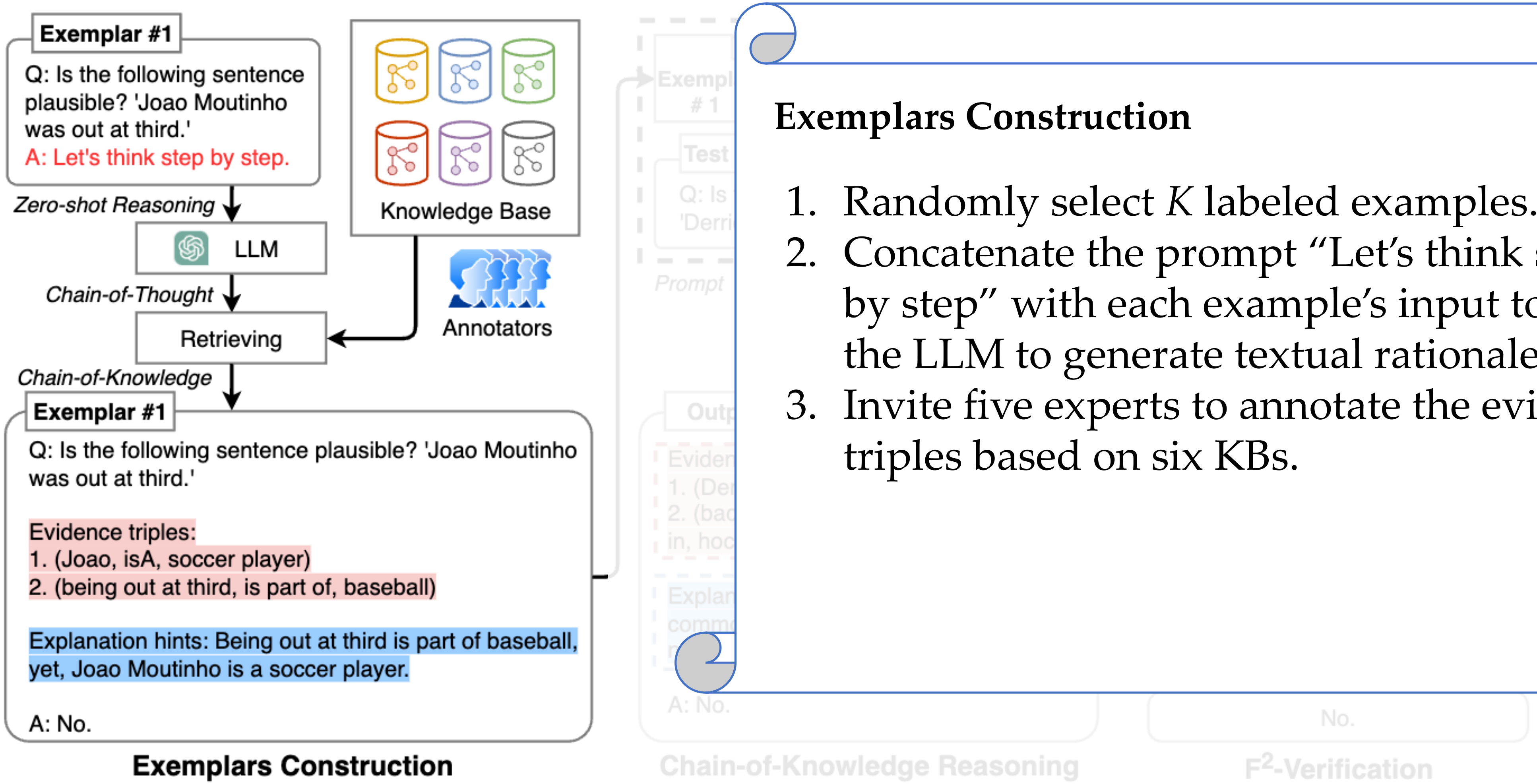
Faithfulness and Factuality

F^2-Verification:

Merge faithfulness and factuality score

$$s(\widetilde{T}, \widetilde{H} | \widetilde{X}, \mathcal{G}) = \gamma \times s_u(\widetilde{T}, \widetilde{H} | \widetilde{X}) + (1 - \gamma) \times s_v(\widetilde{T} | \widetilde{X}, \mathcal{G})$$

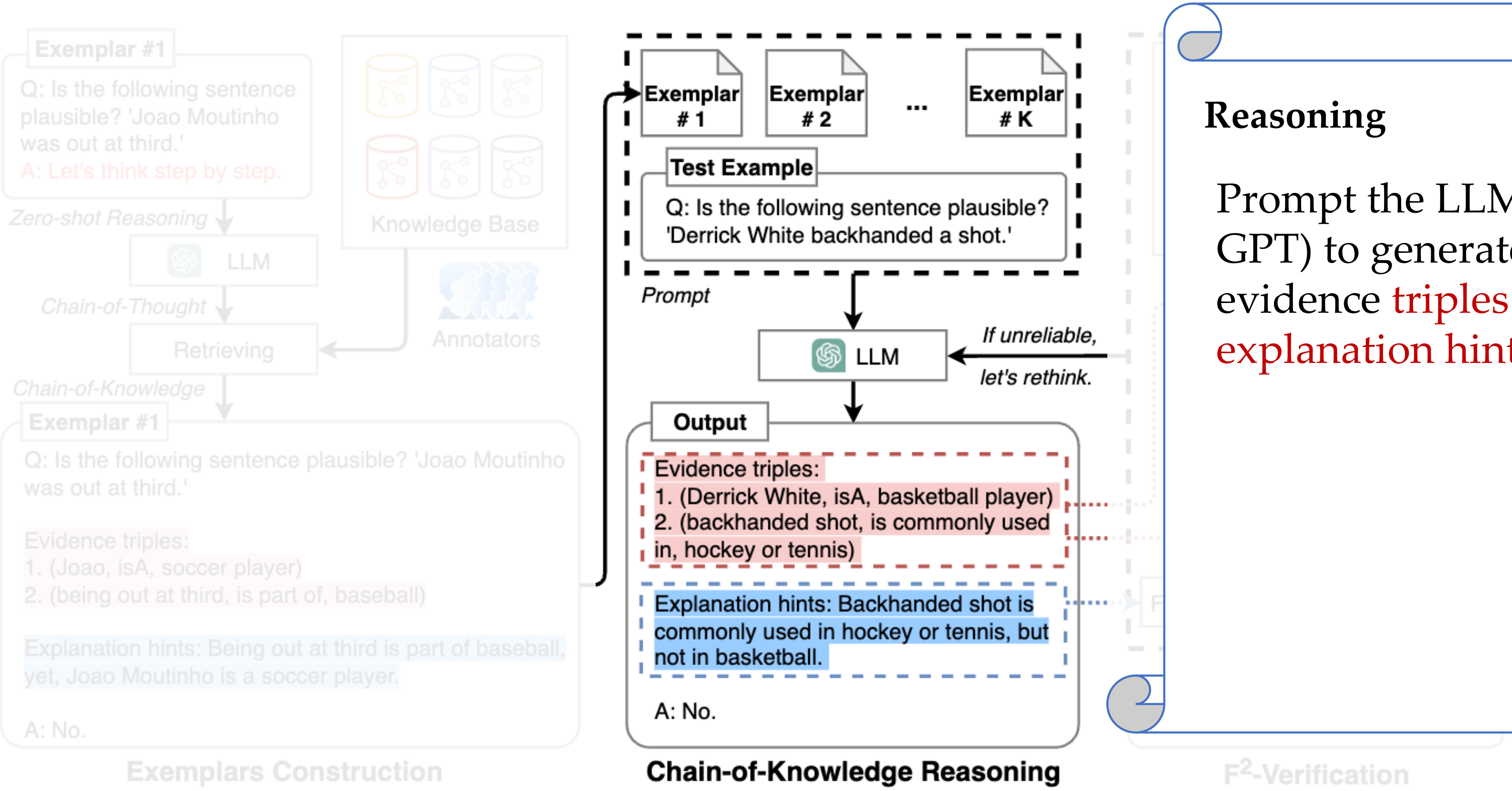
Overview Framework


Exemplar #1
Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'
A: Let's think step by step.
Zero-shot Reasoning
LLM
Chain-of-Thought
Retrieving
Chain-of-Knowledge
Knowledge Base
Annotators
Exemplar #1
Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'
Evidence triples:
1. (Joao, isA, soccer player)
2. (being out at third, is part of, baseball)
Explanation hints: Being out at third is part of baseball, yet, Joao Moutinho is a soccer player.
A: No.
Exemplars Construction
Prompt
A: No.
No.
Chain-of-Knowledge Reasoning
F2-Verification

Exemplars Construction

1. Randomly select K labeled examples.
2. Concatenate the prompt "Let's think step by step" with each example's input to elicit the LLM to generate textual rationale.
3. Invite five experts to annotate the evidence triples based on six KBs.

Overview Framework



Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'
A: Let's think step by step.

Q: Is the following sentence plausible? 'Joao Moutinho was out at third.'

Evidence triples:
1. (Joao, isA, soccer player)
2. (being out at third, is part of, baseball)

Explanation hints: Being out at third is part of baseball, yet, Joao Moutinho is a soccer player.

A: No.

Q: Is the following sentence plausible? 'Derrick White backhanded a shot.'

Evidence triples:
1. (Derrick White, isA, basketball player)
2. (backhanded shot, is commonly used in, hockey or tennis)

Explanation hints: Backhanded shot is commonly used in hockey or tennis, but not in basketball.

A: No.

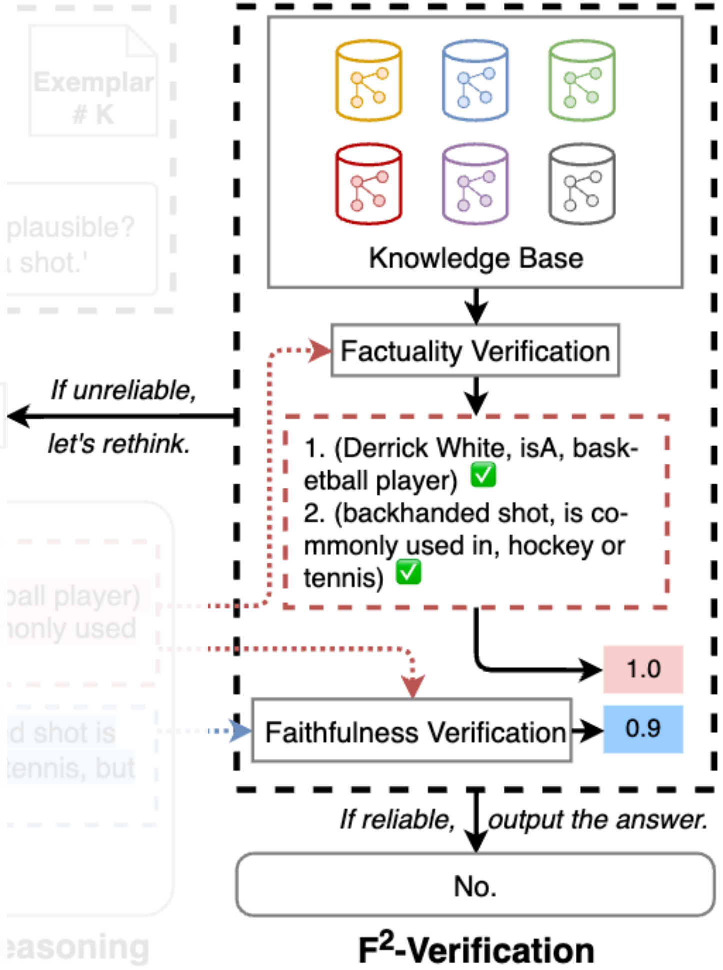
Reasoning

Prompt the LLM (e.g., GPT) to generate both evidence triples and explanation hints.

Overview Framework

F^2-Verification

1. Calculate the faithfulness score and factuality score.
2. Merge two scores to form reliable score.
3. If the reliable score is lower than a threshold, then perform *Rethinking Process*.

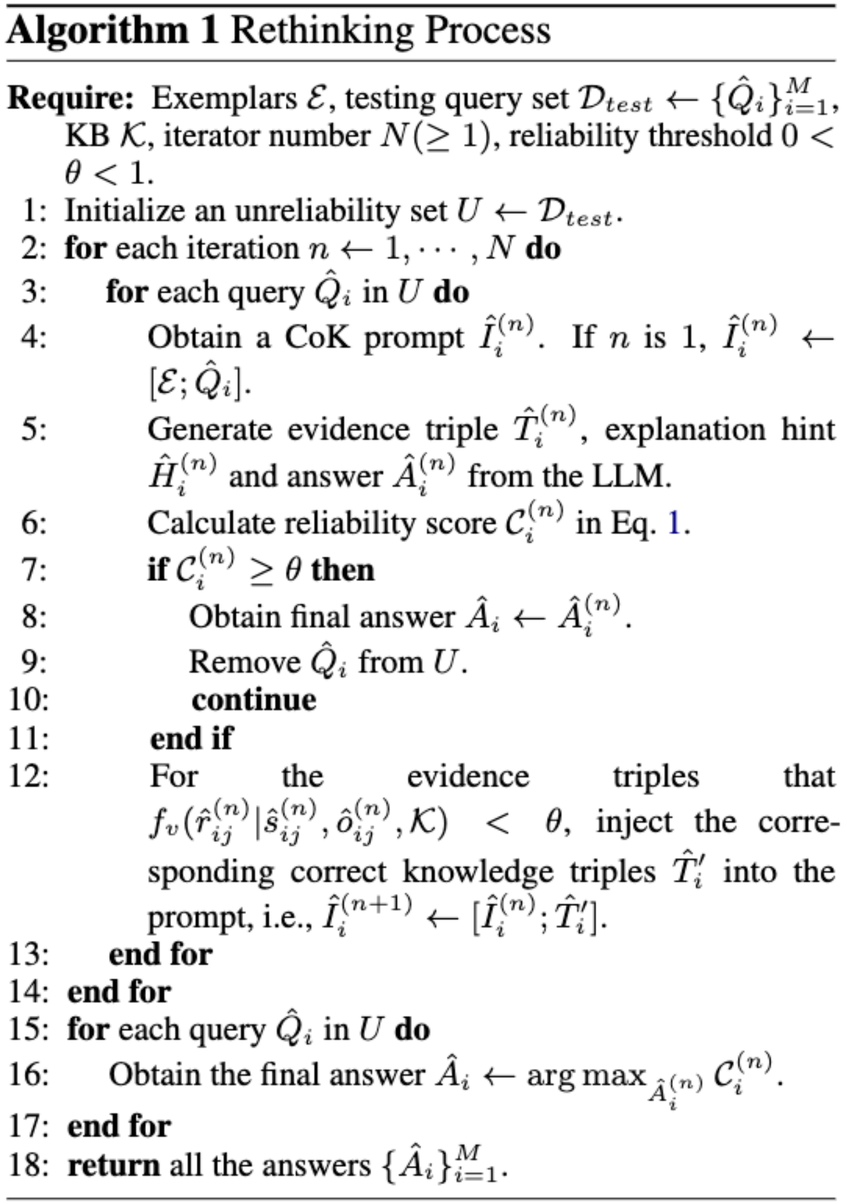


Knowledge Base
Factuality Verification
1. (Derrick White, isA, basketball player) ✅
2. (backhanded shot, is commonly used in, hockey or tennis) ✅
1.0
Faithfulness Verification
0.9
If unreliable,
let's rethink.
If reliable, output the answer.
No.

F^2-Verification

Rethinking

- F^2-Verification ensures factuality and faithfulness of triples and explanations.
- Rethinking algorithm iteratively refines answers based on reliability
- Queries are repeatedly evaluated and improved by injecting correct knowledge.
- Dynamically boosts the performance of LLMs by refining the reasoning chains.

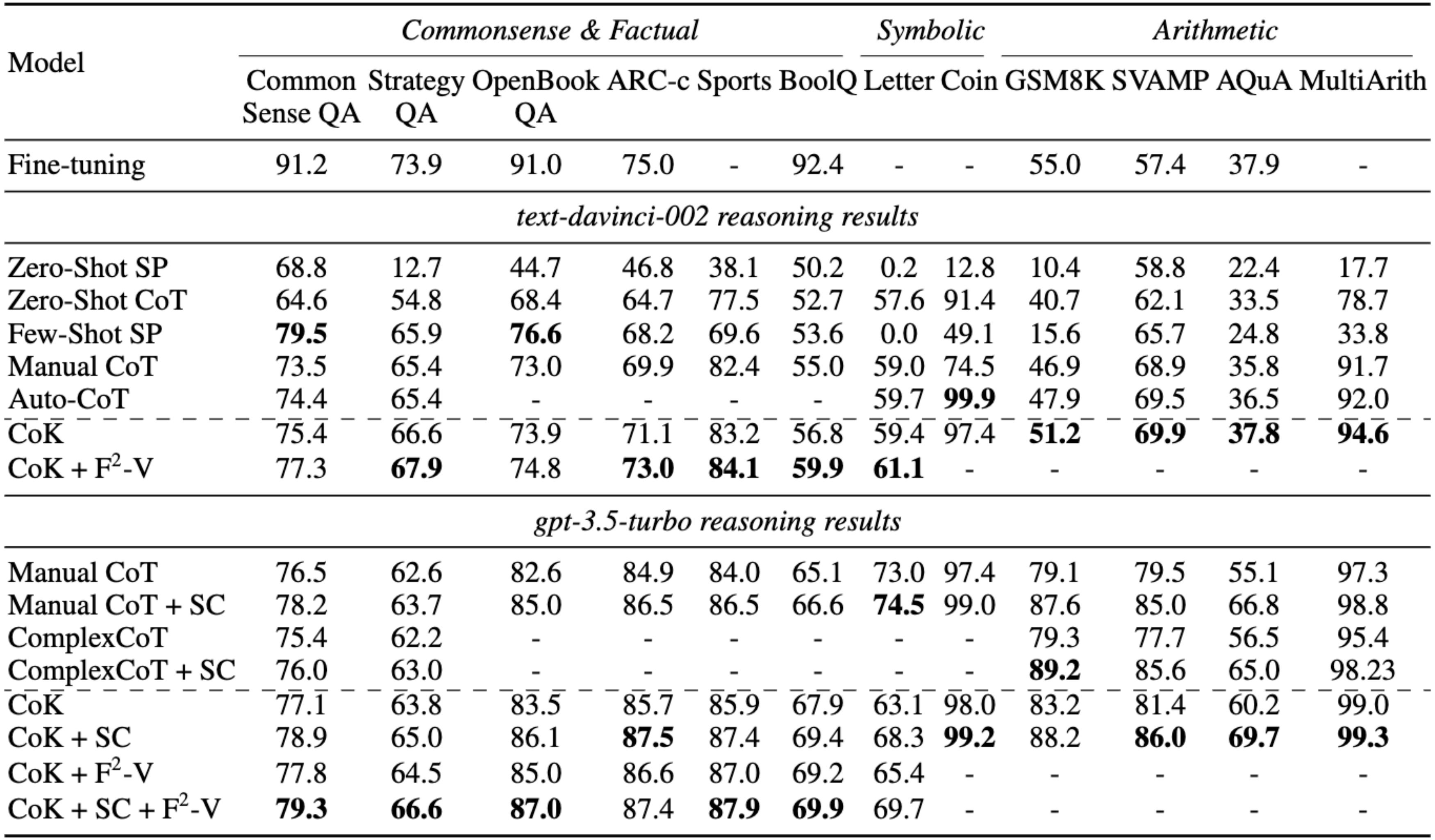


Algorithm 1 Rethinking Process

```
Require: Exemplars $\mathcal{E}$, testing query set $\mathcal{D}_{test} \leftarrow \{\hat{Q}_i\}_{i=1}^{M}$,
    KB $\mathcal{K}$, iterator number $N(\geq 1)$, reliability threshold $0 <$
    $\theta < 1$.
Initialize an unreliability set $U \leftarrow \mathcal{D}_{test}$.
for each iteration $n \leftarrow 1, \cdots, N$ do
    for each query $\hat{Q}_i$ in $U$ do
        Obtain a CoK prompt $\hat{I}_i^{(n)}$. If $n$ is 1, $\hat{I}_i^{(n)} \leftarrow$
        $[\mathcal{E}; \hat{Q}_i]$.
        Generate evidence triple $\hat{T}_i^{(n)}$, explanation hint
        $\hat{H}_i^{(n)}$ and answer $\hat{A}_i^{(n)}$ from the LLM.
        Calculate reliability score $\mathcal{C}_i^{(n)}$ in Eq. 1.
        if $\mathcal{C}_i^{(n)} \geq \theta$ then
            Obtain final answer $\hat{A}_i \leftarrow \hat{A}_i^{(n)}$.
            Remove $\hat{Q}_i$ from $U$.
            continue
        end if
        For the evidence triples that
        $f_v(\hat{r}_{ij}^{(n)} | \hat{s}_{ij}^{(n)}, \hat{o}_{ij}^{(n)}, \mathcal{K}) < \theta$, inject the corre-
        sponding correct knowledge triples $\hat{T}_i'$ into the
        prompt, i.e., $\hat{I}_i^{(n+1)} \leftarrow [\hat{I}_i^{(n)}; \hat{T}_i']$.
    end for
end for
for each query $\hat{Q}_i$ in $U$ do
    Obtain the final answer $\hat{A}_i \leftarrow \arg\max_{\hat{A}_i^{(n)}} \mathcal{C}_i^{(n)}$.
end for
return all the answers $\{\hat{A}_i\}_{i=1}^{M}$.
```

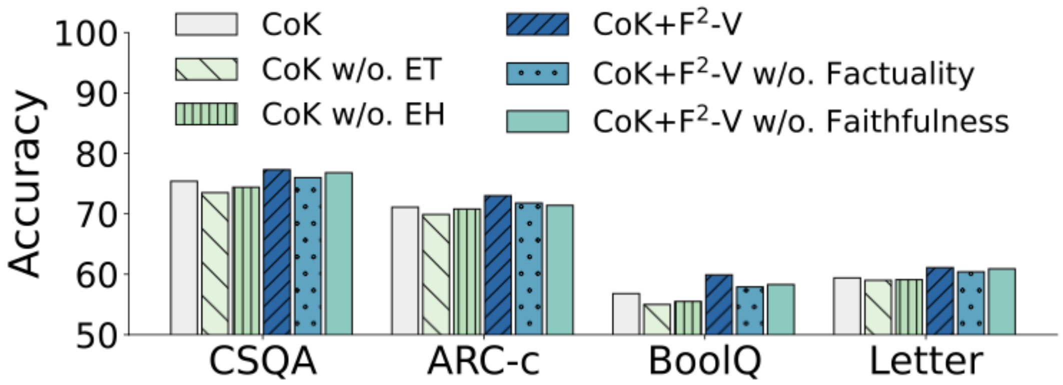
Empirical Evaluations

Overall



Model	*Commonsense & Factual*						*Symbolic*		*Arithmetic*			
	Common Sense QA	Strategy QA	OpenBook QA	ARC-c	Sports	BoolQ	Letter	Coin	GSM8K	SVAMP	AQuA	MultiArith
Fine-tuning	91.2	73.9	91.0	75.0	-	92.4	-	-	55.0	57.4	37.9	-
text-davinci-002 reasoning results												
Zero-Shot SP	68.8	12.7	44.7	46.8	38.1	50.2	0.2	12.8	10.4	58.8	22.4	17.7
Zero-Shot CoT	64.6	54.8	68.4	64.7	77.5	52.7	57.6	91.4	40.7	62.1	33.5	78.7
Few-Shot SP	**79.5**	65.9	**76.6**	68.2	69.6	53.6	0.0	49.1	15.6	65.7	24.8	33.8
Manual CoT	73.5	65.4	73.0	69.9	82.4	55.0	59.0	74.5	46.9	68.9	35.8	91.7
Auto-CoT	74.4	65.4	-	-	-	-	59.7	**99.9**	47.9	69.5	36.5	92.0
CoK	75.4	66.6	73.9	71.1	83.2	56.8	59.4	97.4	**51.2**	**69.9**	**37.8**	**94.6**
CoK + F^2-V	77.3	**67.9**	74.8	**73.0**	**84.1**	**59.9**	**61.1**	-	-	-	-	-
gpt-3.5-turbo reasoning results												
Manual CoT	76.5	62.6	82.6	84.9	84.0	65.1	73.0	97.4	79.1	79.5	55.1	97.3
Manual CoT + SC	78.2	63.7	85.0	86.5	86.5	66.6	**74.5**	99.0	87.6	85.0	66.8	98.8
ComplexCoT	75.4	62.2	-	-	-	-	-	-	79.3	77.7	56.5	95.4
ComplexCoT + SC	76.0	63.0	-	-	-	-	-	-	**89.2**	85.6	65.0	98.23
CoK	77.1	63.8	83.5	85.7	85.9	67.9	63.1	98.0	83.2	81.4	60.2	99.0
CoK + SC	78.9	65.0	86.1	**87.5**	87.4	69.4	68.3	**99.2**	88.2	**86.0**	**69.7**	**99.3**
CoK + F^2-V	77.8	64.5	85.0	86.6	87.0	69.2	65.4	-	-	-	-	-
CoK + SC + F^2-V	**79.3**	**66.6**	**87.0**	87.4	**87.9**	**69.9**	69.7	-	-	-	-	-

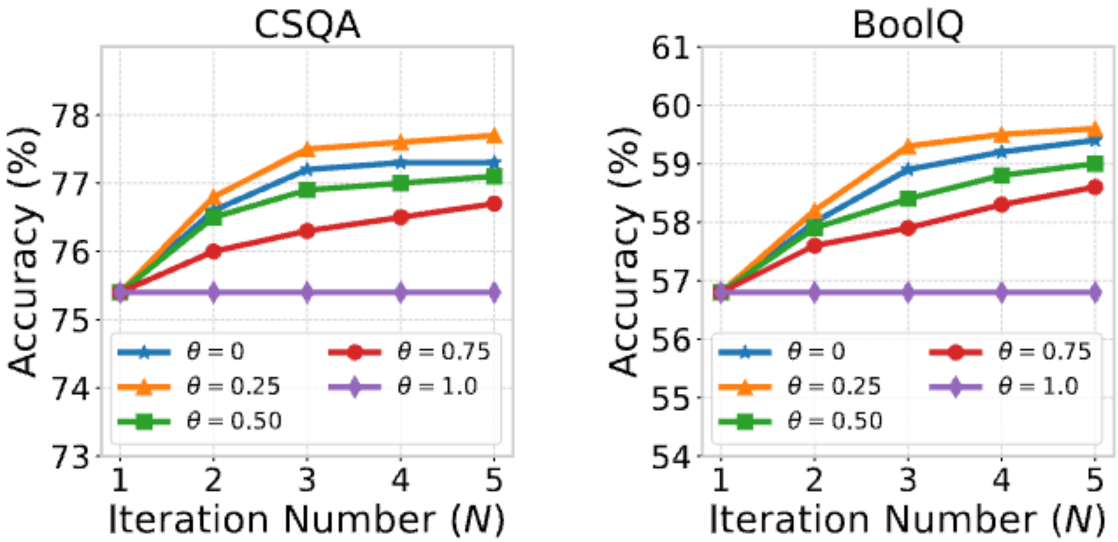
Ablations



- Observations:
  - □ Performance drops when removing any component.
  - □ the variant without explicit evidence triples (CoK w/o. ET) performs worse than without explanation hints (CoK w/o. EH)
  - □Both triples and explanation hints guide LLMs to verify reasoning chains

Takeaway: All components are crucial, with explicit evidence triples being the most significant for performance.

Rethinking Effectiveness



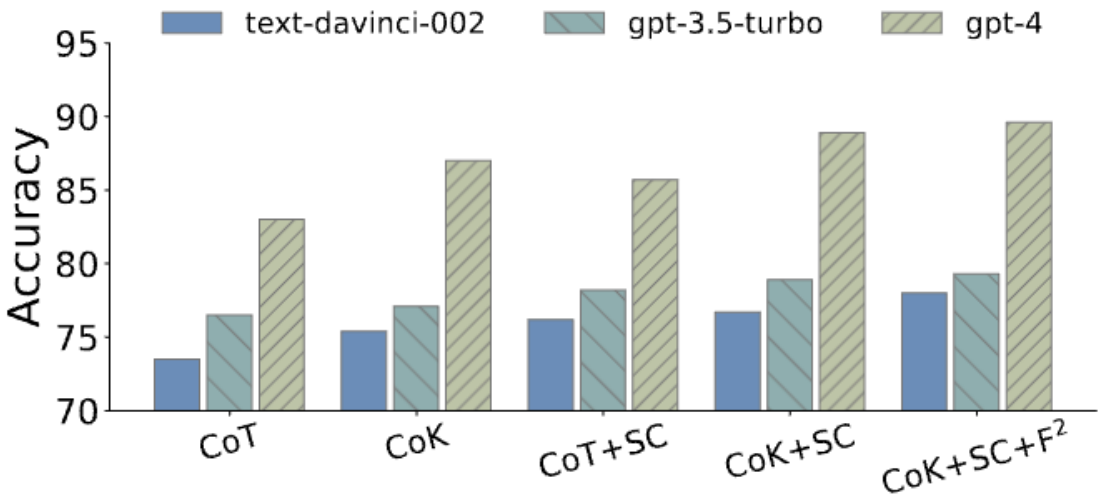
■ Observations:

- □ Accuracy significantly increases during the first 3 iterations when the LLM rethinks step-by-step

- □ Performance may drop due to over-injection of irrelevant or inconsistent information.

Takeaway: Step-by-step rethinking improves acc. but requires careful threshold management to avoid over-injection.

Different LLMs



- Observations:
  - Using GPT-4 to evaluate CSQA and GSM8K, showing CoK and CoK+SC work well.
  - From avg. CoK demonstrates its adaptability

Takeaway: Improved performance across different LLMs, demonstrating versatility and effectiveness.

Conclusions

Conclusions

- CoK, a method to elicit LLMs for generating explicit structured rationale.
- Introduce faithfulness and factuality evaluation for enhanced rationale correctness.
- Propose a rethinking algorithm to reduce the hallucination through an iteration process with external KBs.
- Achieve good performance on multiple reasoning tasks.

Thanks For Listening !

💻 wjn1996/Chain-of-Knowledge 🐦 @qiushi_sun

✉ lygwjn@gmail.com or qiushisun@connect.hku.hk