Maeutic Prompting

Logically Consistent Reasoning with Recursive Explanations

Sean Welleck | 09.19.2022

Led by:
Jaehun Jung

Lianhui Qin

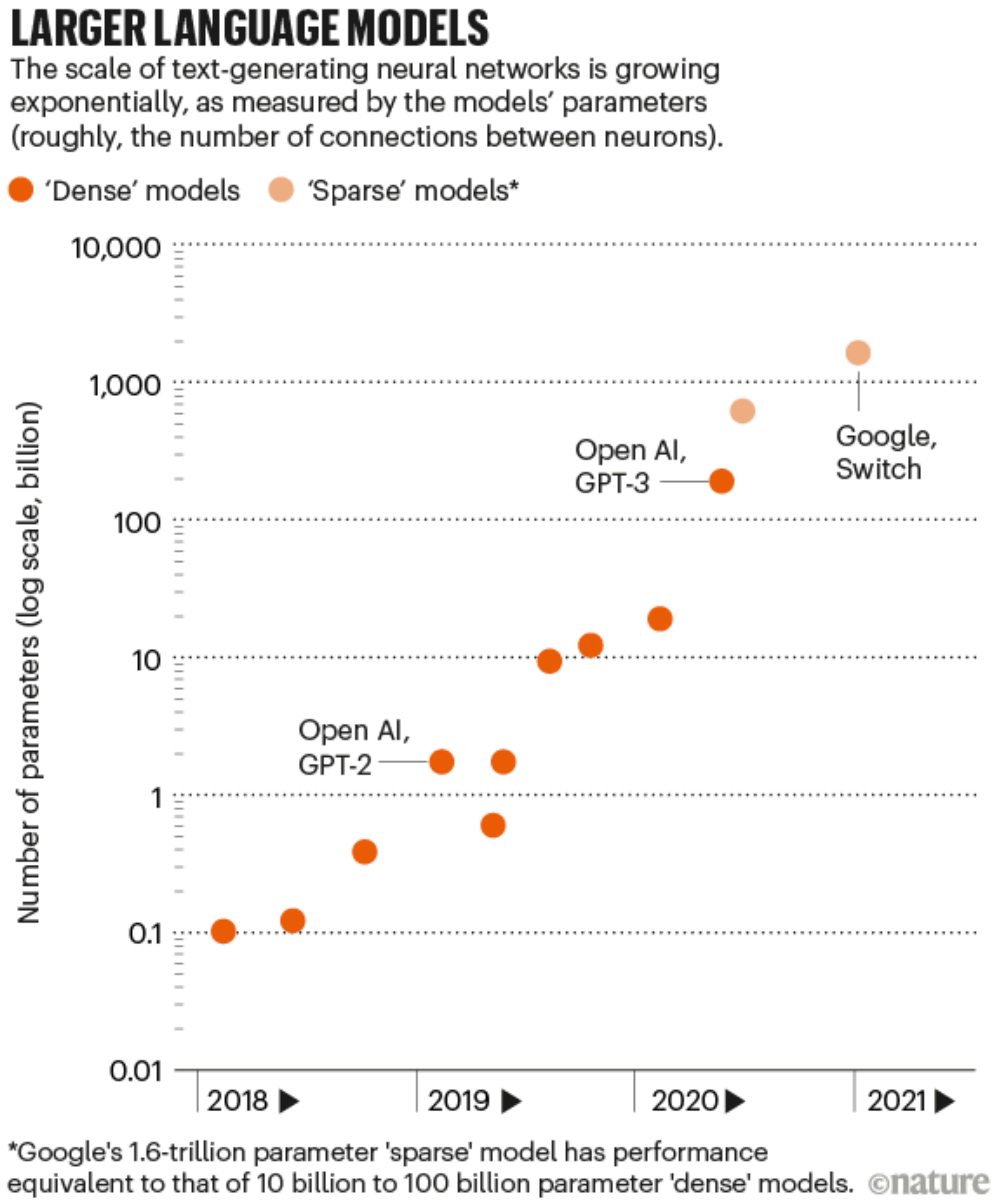
Faeze
Brahman

Chandra
Bhagavatula

Ronan
Le Bras

Yejin Choi

Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations
https://arxiv.org/abs/2205.11822
Under Review



[Peters et al. ’18 , Radford et al. ’19, Brown et al. ’20, ….]

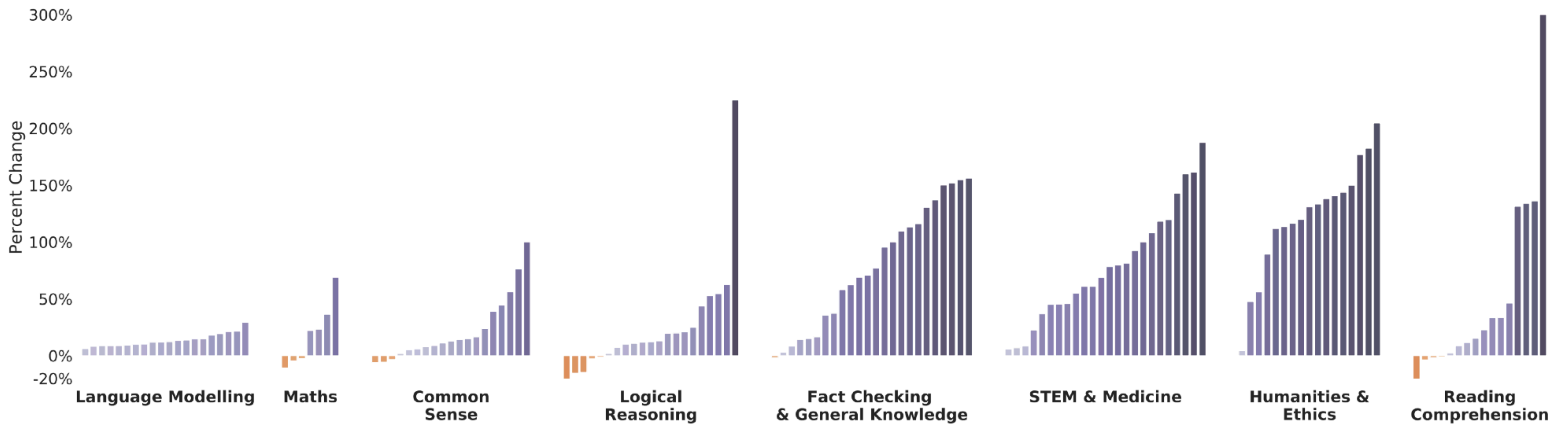


Figure 4 | **280B vs best performance up to 7.1B** across different tasks. We compare the performance

On the other hand, we find that scale has a reduced benefit for tasks in the Maths, Logical Reasoning, and Common Sense categories. Our results suggest that for certain flavours of mathematical or logical reasoning tasks, it is unlikely that *scale* alone will lead to performance breakthroughs. In some cases *Gopher* has a lower performance than smaller models– examples of which include **Abstract Algebra** and **Temporal Sequences** from BIG-bench, and **High School Mathematics** from MMLU.

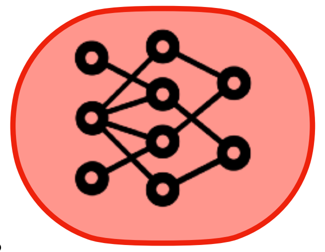
[Rae et al (Deepmind) 2022, Scaling Language Models: Methods, Analysis & Insights from Training Gopher]

Claim Verification

Claim: One can drive La Jolla to New York City in less than two hours. **FALSE**

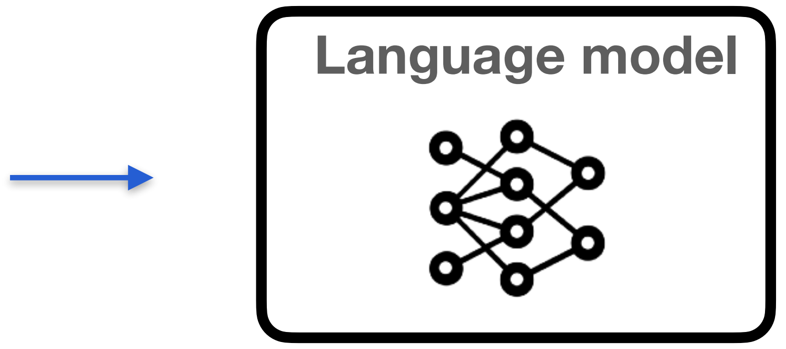
Claim: Harry Potter can teach classes on how to fly on a broomstick. **TRUE**

Claim: One is a number that comes ***after*** zero. GPT-3 175B **TRUE**



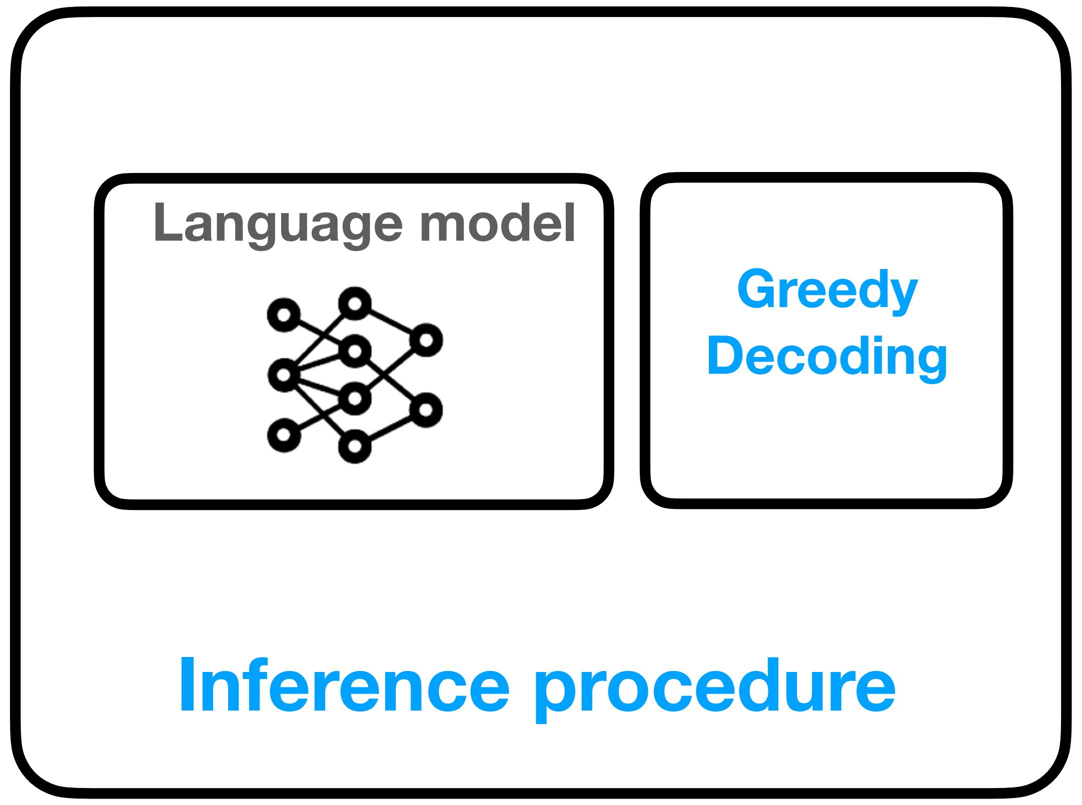
Claim: One is a number that comes ***before*** zero. **TRUE**

Claim: One is a number that comes ***before*** zero.



Claim: One is a number that comes ***before*** zero.

x



True

y

$$y = \text{argmax}_y \; p(y \mid x)$$

Better inference procedure?

Explanation-based prompting & inference

- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)

Diagram: Wang et al 2022, Self-Consistency Improves Chain of Thought Reasoning in Language Models

Explanation-based prompting & inference

- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)
  - $y \sim p(y \mid z, x)$ answer given z

Diagram: Wang et al 2022, Self-Consistency Improves Chain of Thought Reasoning in Language Models

Explanation-based prompting & inference

- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)
  - $y \sim p(y \mid z, x)$ answer given z
- Some LMs can be *prompted* to generate z [Wei et al 2022]

Diagram: Wang et al 2022, Self-Consistency Improves Chain of Thought Reasoning in Language Models

Explanation-based prompting & inference

- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)
  - $y \sim p(y \mid z, x)$ answer given z
- Some LMs can be *prompted* to generate *z* [Wei et al 2022]
- Variations, e.g. sample multiple z's and aggregate y's [Wang et al 2022]

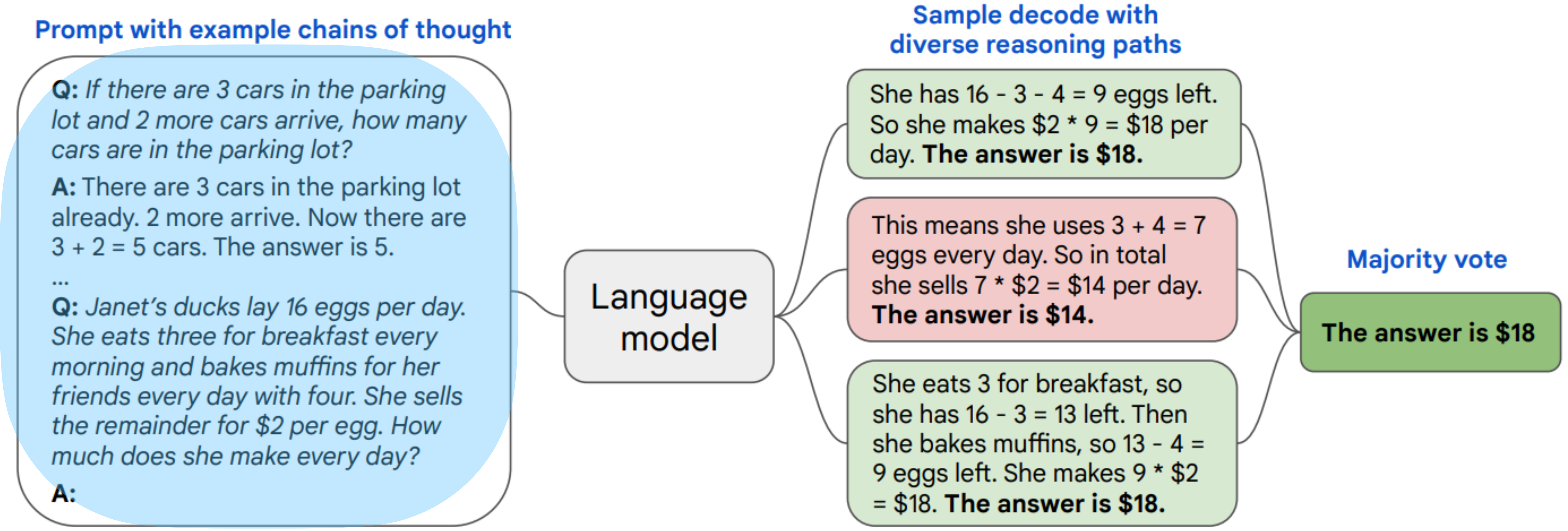
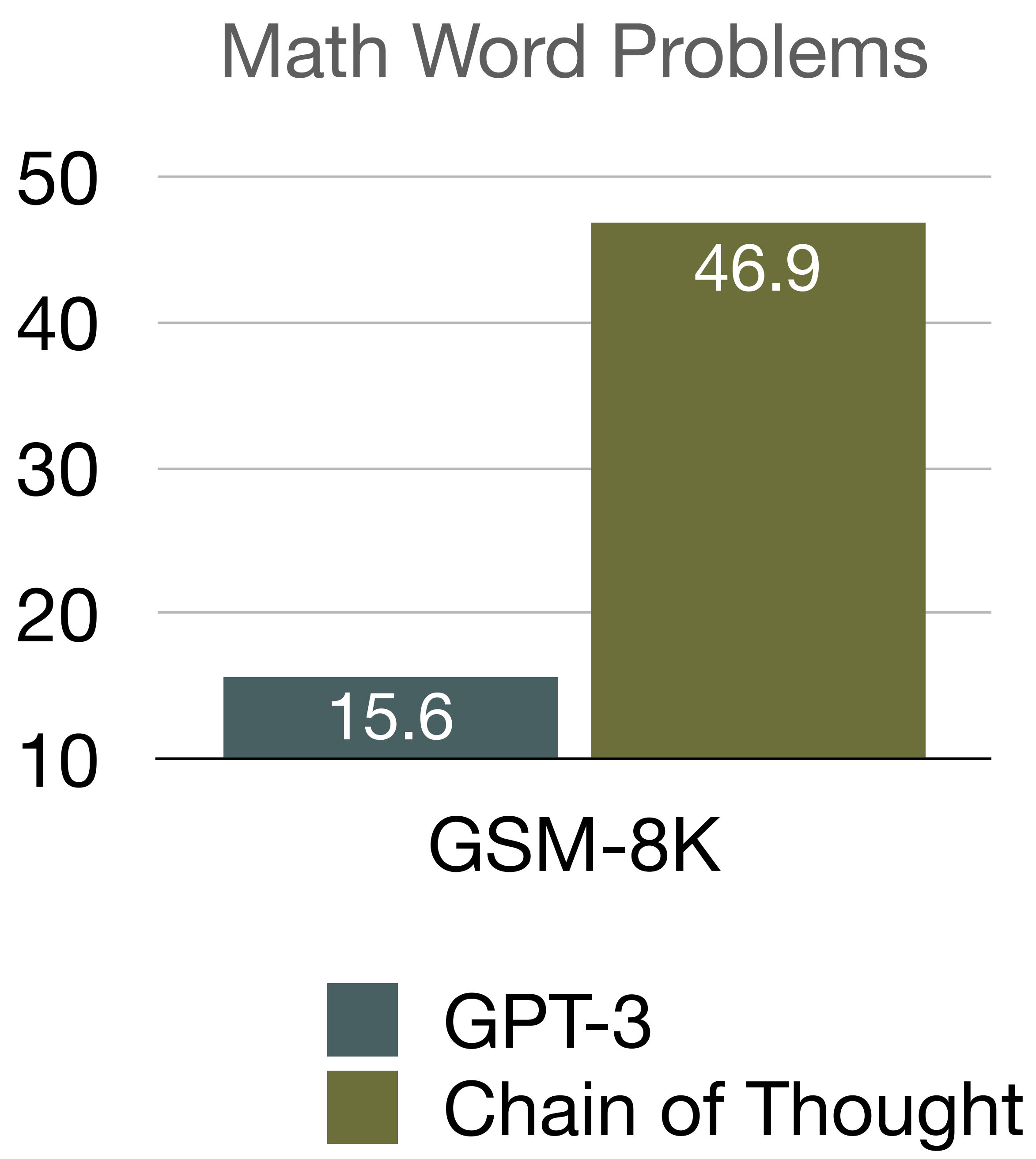


Diagram: Wang et al 2022, Self-Consistency Improves Chain of Thought Reasoning in Language Models

Explanation-based prompting & inference

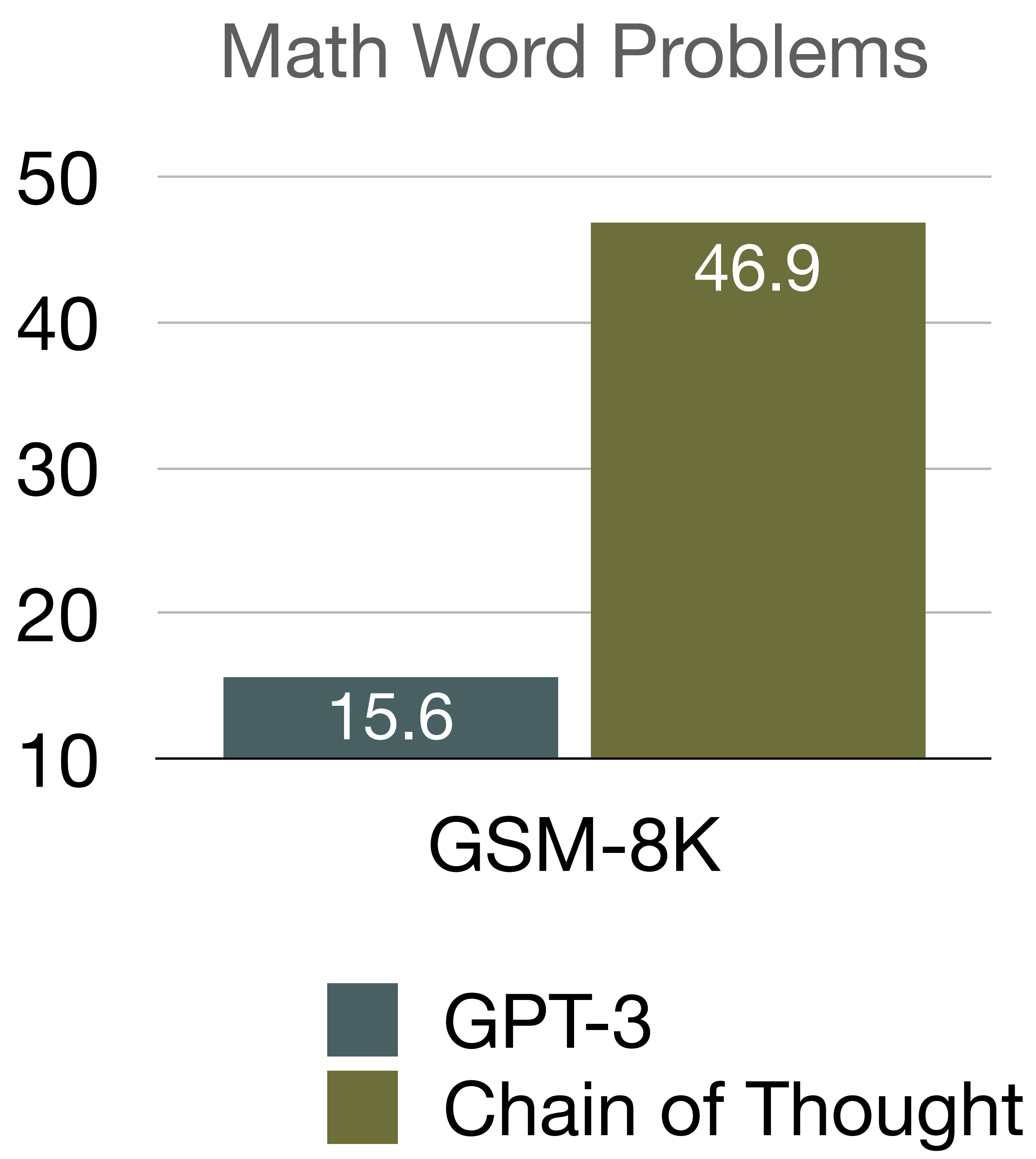
- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)
  - $y \sim p(y \mid z, x)$ answer given z



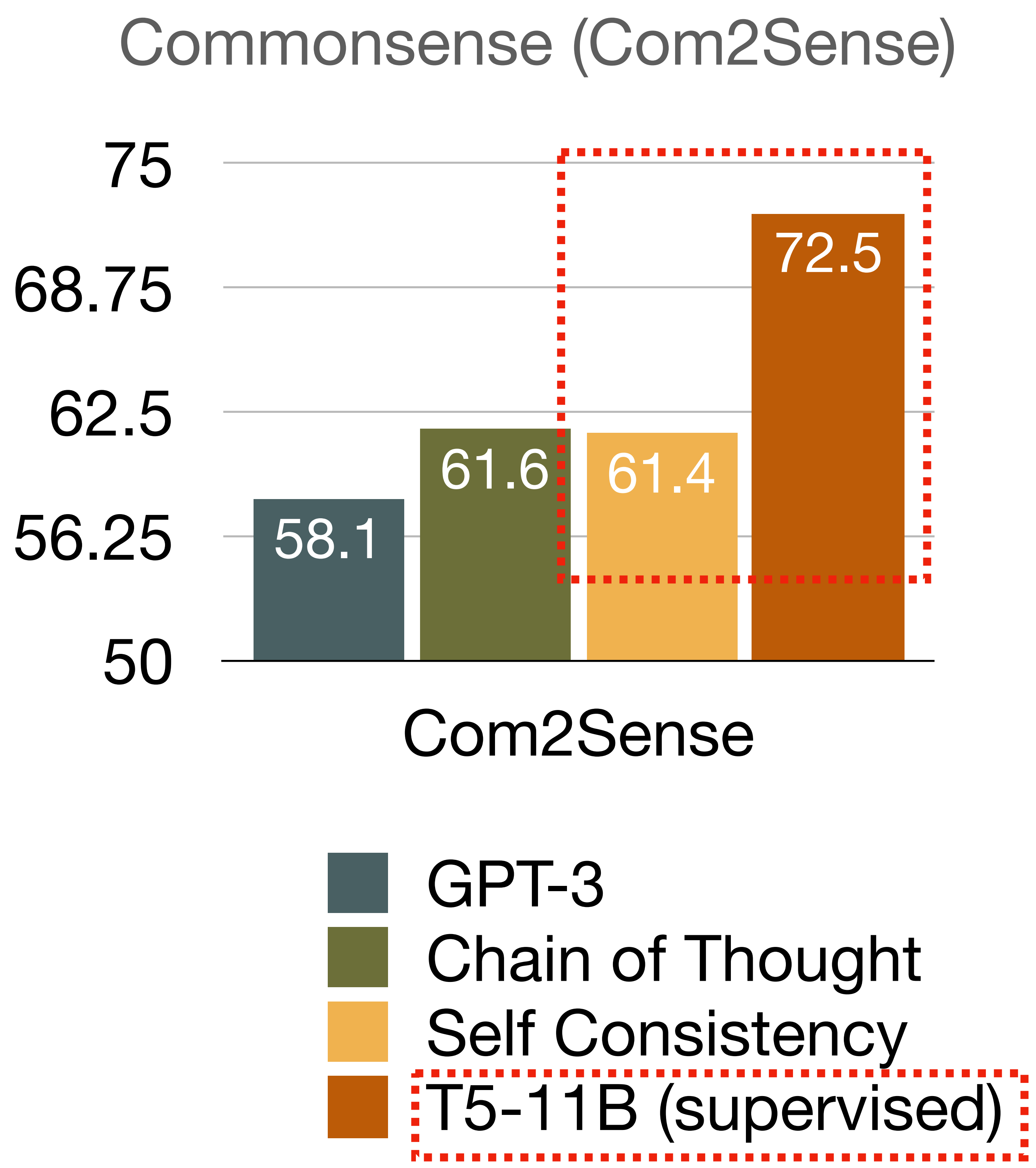
GSM8k result: Wei et al 2022

Explanation-based prompting & inference

- Factor generation into two stages:
  - $z \sim p(z \mid x; D)$ intermediate sequence z (explanation/rationale/chain of thought/reasoning path/…)
  - $y \sim p(y \mid z, x)$ answer given z



GSM8k result: Wei et al 2022



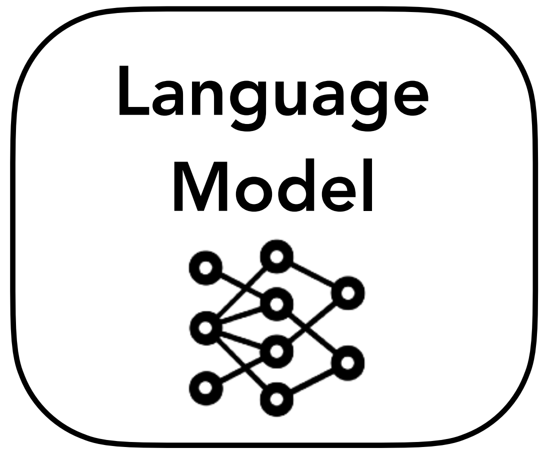
Our experiments

Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.



GPT3 175B (text-davinci-001)

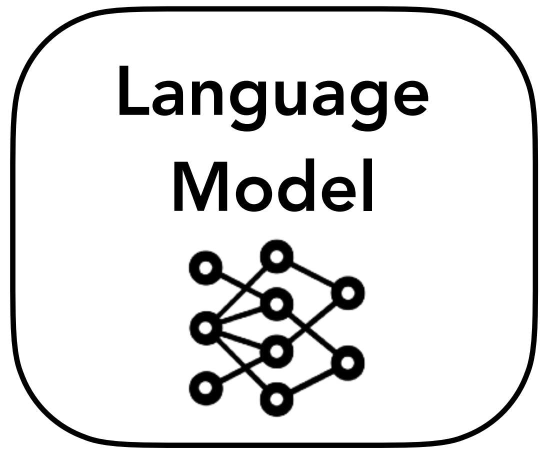
See Also:
"The Unreliability of Explanations in Few-Shot In-Context Learning"
Xi Ye, Greg Durrett

Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.



z y

Smoke is a result of fire. Therefore, the statement is False.

GPT3 175B (text-davinci-001)

See Also:
"The Unreliability of Explanations in Few-Shot In-Context Learning"
Xi Ye, Greg Durrett

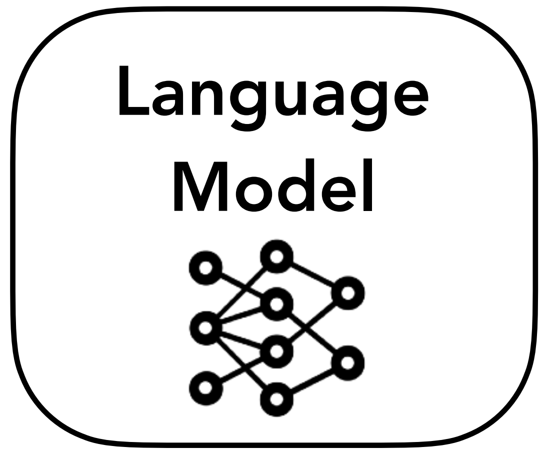
Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.

¬source



z

Smoke is a result of fire.

y

Therefore, the statement is False.

GPT3 175B (text-davinci-001)

See Also:
"The Unreliability of Explanations in Few-Shot In-Context Learning"
Xi Ye, Greg Durrett

Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.

¬source

Language Model

z

Smoke is a result of fire.

y

Therefore, the statement is False.

`result`

GPT3 175B (text-davinci-001)

See Also:
“The Unreliability of Explanations in Few-Shot In-Context Learning”
Xi Ye, Greg Durrett

Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.

¬source

Language Model

z y

Smoke is a result of fire. Therefore, the statement is False.

`result`

"Common sense"

`result` $\Longrightarrow$ `¬source`

$\therefore$ `¬source`

GPT3 175B (text-davinci-001)

See Also:
"The Unreliability of Explanations in Few-Shot In-Context Learning"
Xi Ye, Greg Durrett

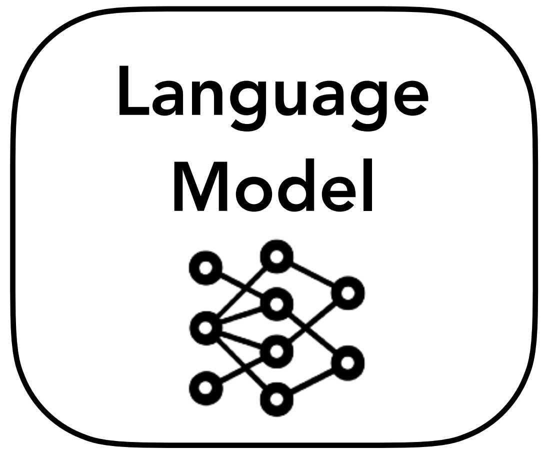
Unreliability of explanations

1. **Incorrect inference:** Explanation does not logically lead to the inferred answer

x

Claim: Smoke is not the source of fire.

$\neg$source



z y

Smoke is a result of fire. Therefore, the statement is False.

`result`

"Common sense"

`result` $\Longrightarrow$ $\neg$`source`

$\therefore$ $\neg$`source`

Model

$\therefore$ `source`

GPT3 175B (text-davinci-001)

See Also:
"The Unreliability of Explanations in Few-Shot In-Context Learning"
Xi Ye, Greg Durrett

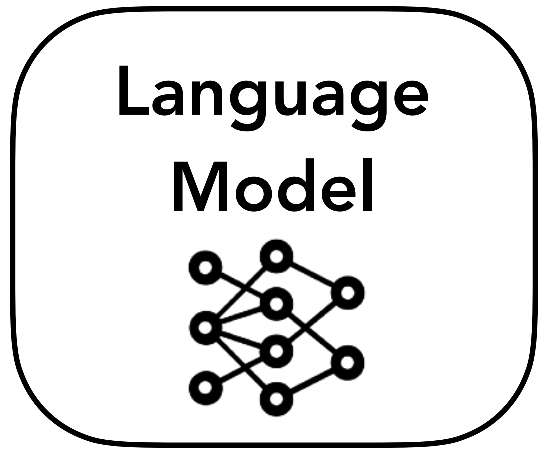
Unreliability of explanations

2. **Logical (non-)integrity**: Same label for a statement and its negation

x

Claim: One is a number that comes **after** zero.

Claim: One is a number that comes **before** zero.



z y

One is … Therefore, the statement is **True**.

One is … Therefore, the statement is **True**.

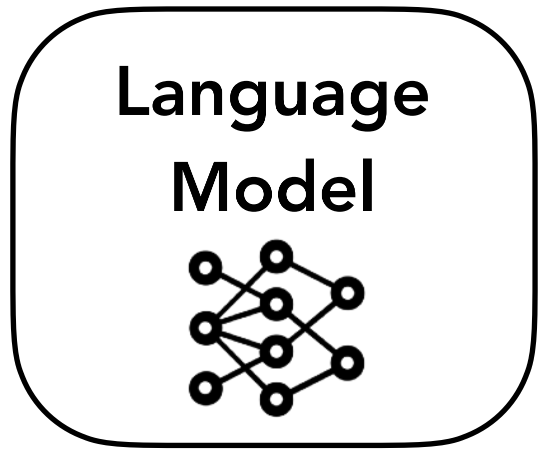
Unreliability of explanations

2. **Logical (non-)integrity**: Same label for a statement and its negation

x

Claim: One is a number that comes **after** zero.

Claim: One is a number that comes **before** zero.



z y

One is … Therefore, the statement is **True**.

One is … Therefore, the statement is **True**.

- Want:

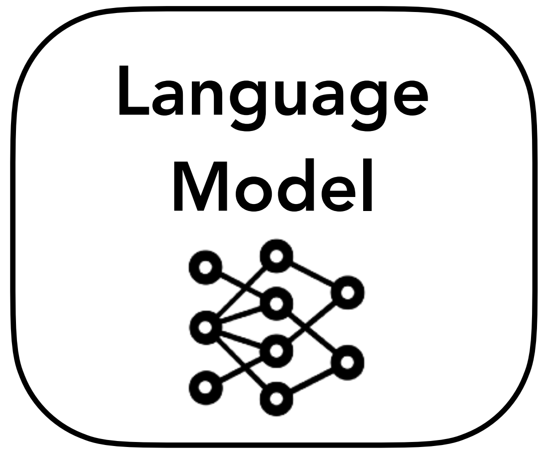
$$\forall p$$

$$f(p) \implies \neg f(\neg p)$$

Unreliability of explanations

3. **Self-contradiction**: model falsifies its own explanation

x		z y
Claim: Butterflies fly with 3 wings.		Butterflies have 4 wings. Therefore, the statement is False.
Claim: Butterflies have 4 wings.		Butterflies have 2 wings on each side of their body. Therefore, the statement is False.



Unreliability of explanations

3. **Self-contradiction**: model falsifies its own explanation

x

Claim: Butterflies fly with 3 wings.

Claim: Butterflies have 4 wings.

Language Model

z

y

Butterflies have 4 wings. Therefore, the statement is False.

Butterflies have 2 wings on each side of their body.
Therefore, the statement is False.

- Want:

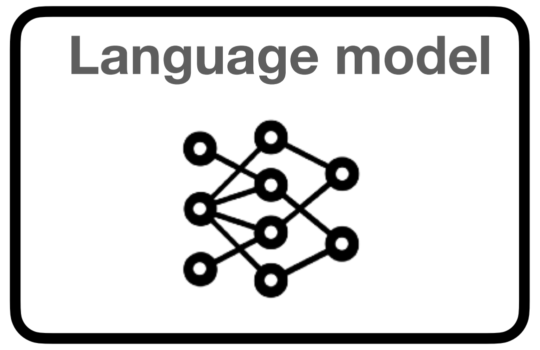
For all model assertions p,
$f(p)$ should evaluate to true

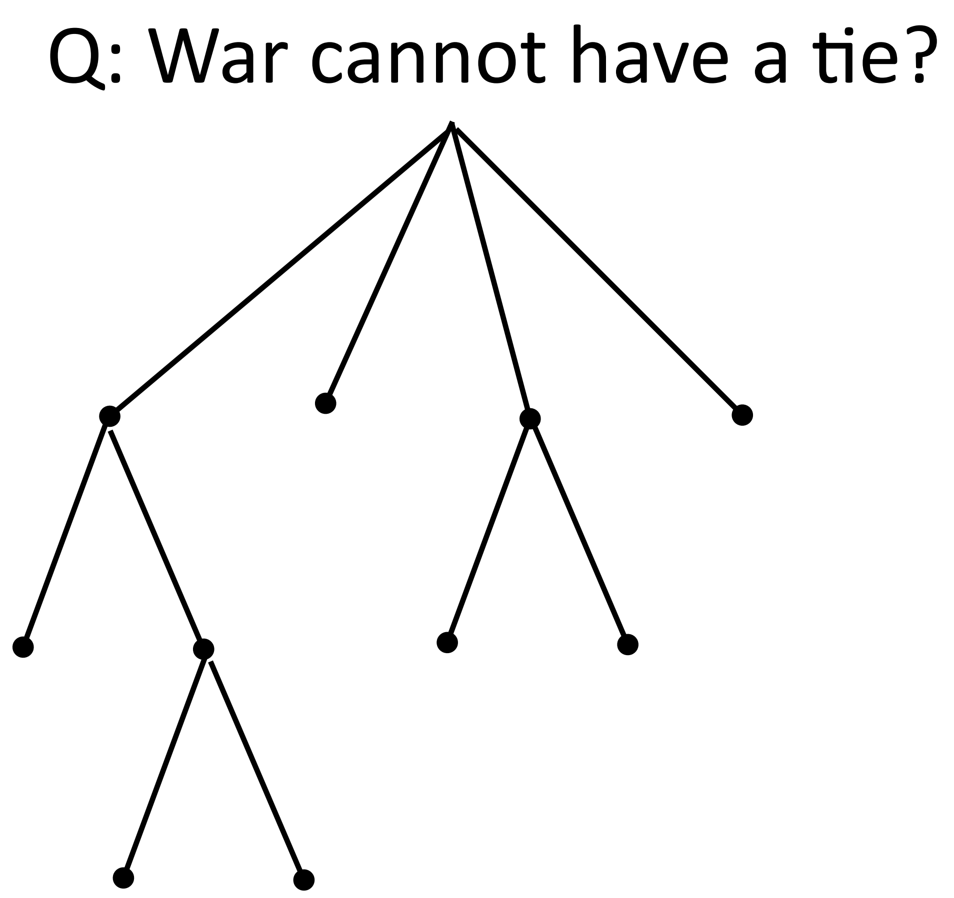
Motivation: inference procedure that accounts for unreliability of explanations

- Take advantage of prompted explanation abilities
  - Account for noisy & contradictory explanations

Motivation: inference procedure that accounts for unreliability of explanations

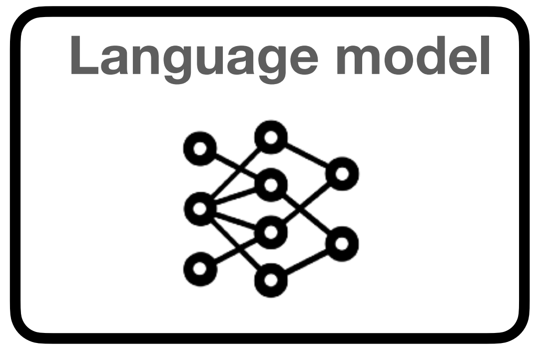
- Take advantage of prompted explanation abilities
  - Account for noisy & contradictory explanations

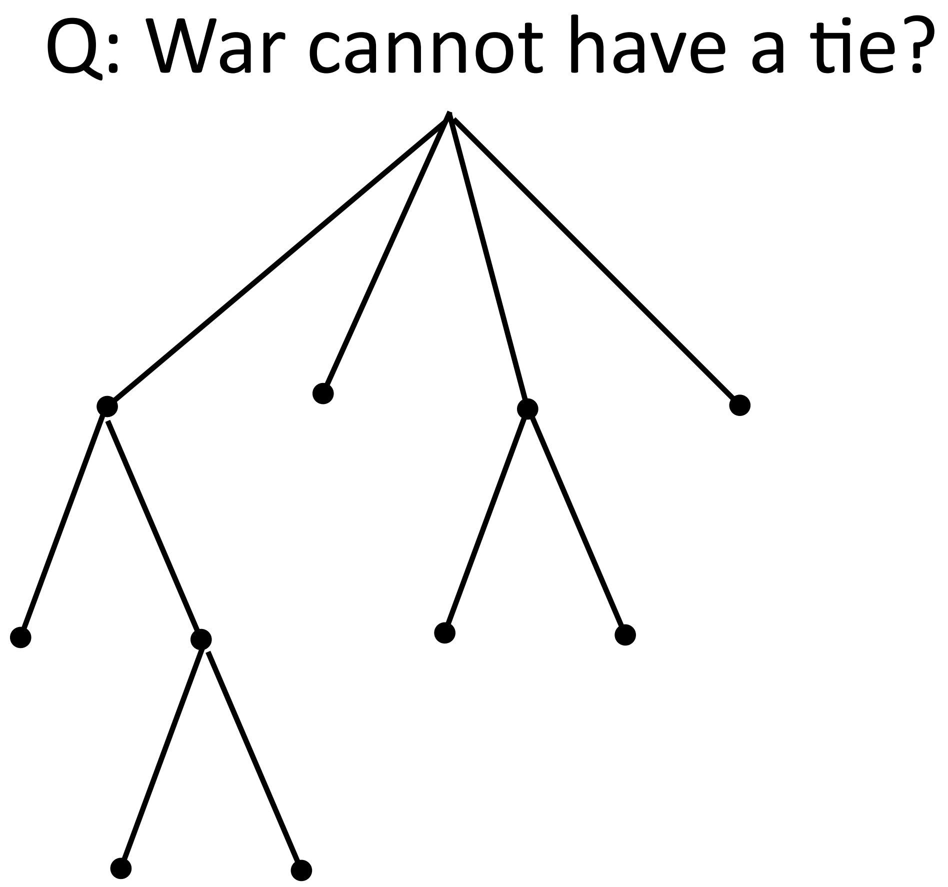




Motivation: inference procedure that accounts for unreliability of explanations

- Take advantage of prompted explanation abilities
  - Account for noisy & contradictory explanations

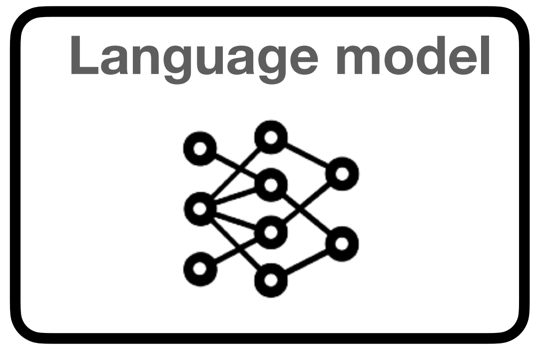


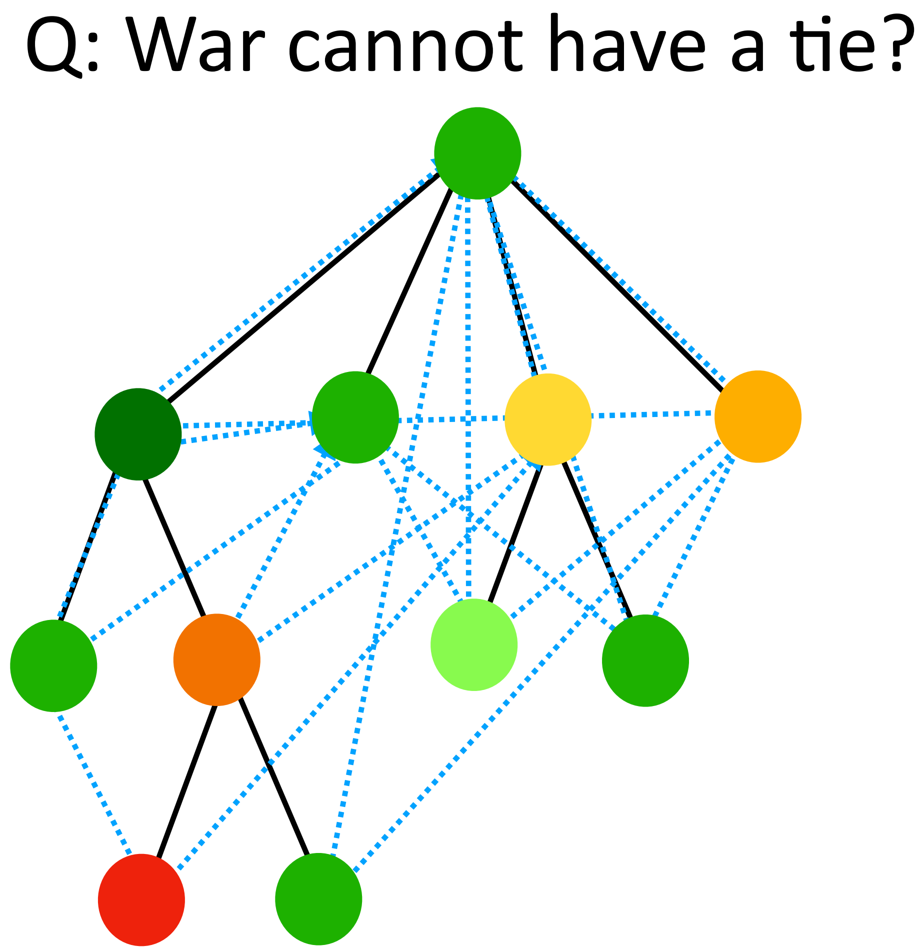


1. Enumerate tree of explanations

Motivation: inference procedure that accounts for unreliability of explanations

- Take advantage of prompted explanation abilities
  - Account for noisy & contradictory explanations

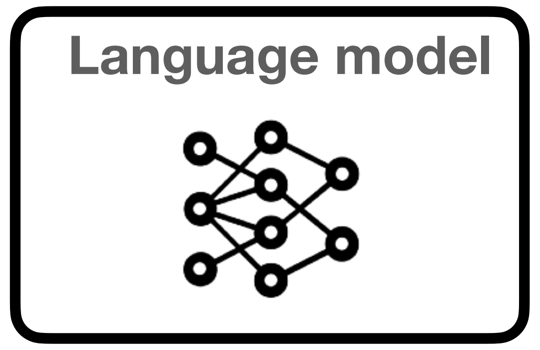


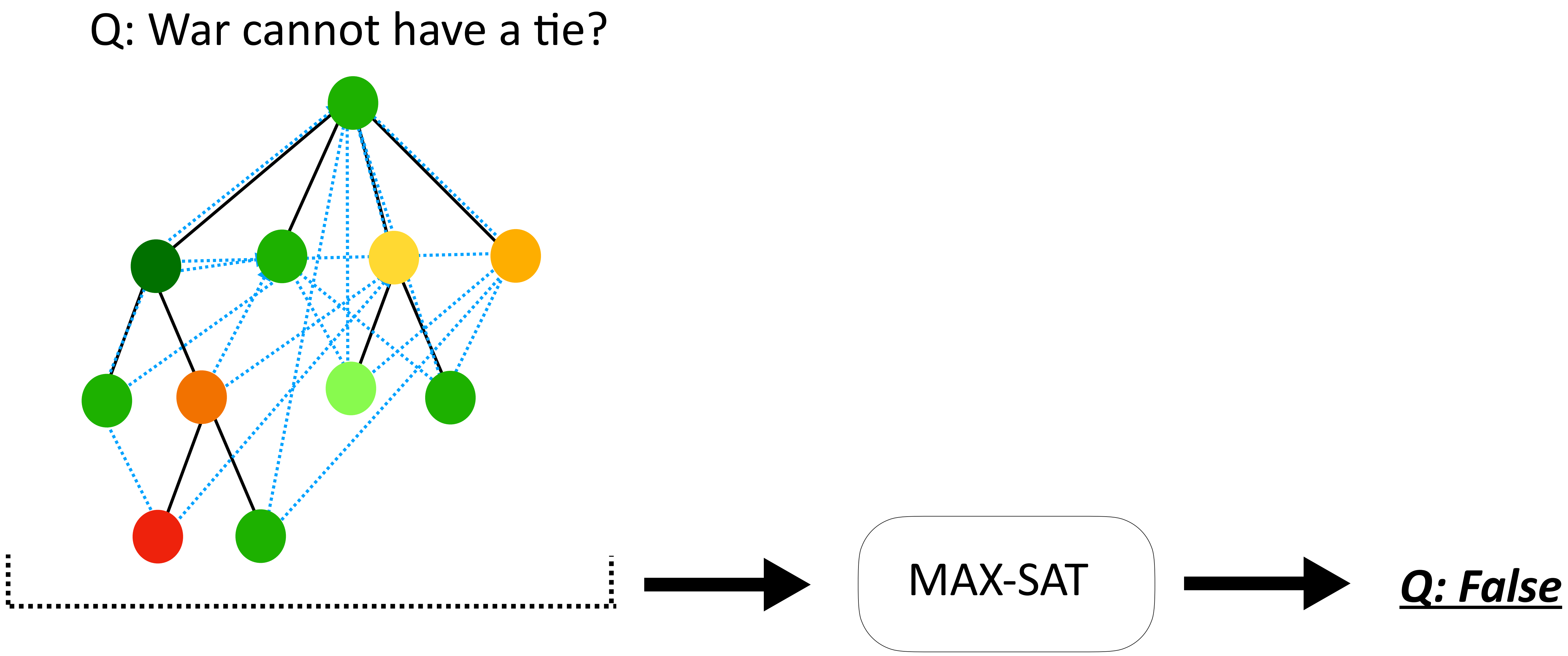


1. Enumerate tree of explanations
2. Score relations in tree

Motivation: inference procedure that accounts for unreliability of explanations

- Take advantage of prompted explanation abilities
    - Account for noisy & contradictory explanations





1. Enumerate tree of explanations

2. Score relations in tree

2. Aggregate scores into a prediction

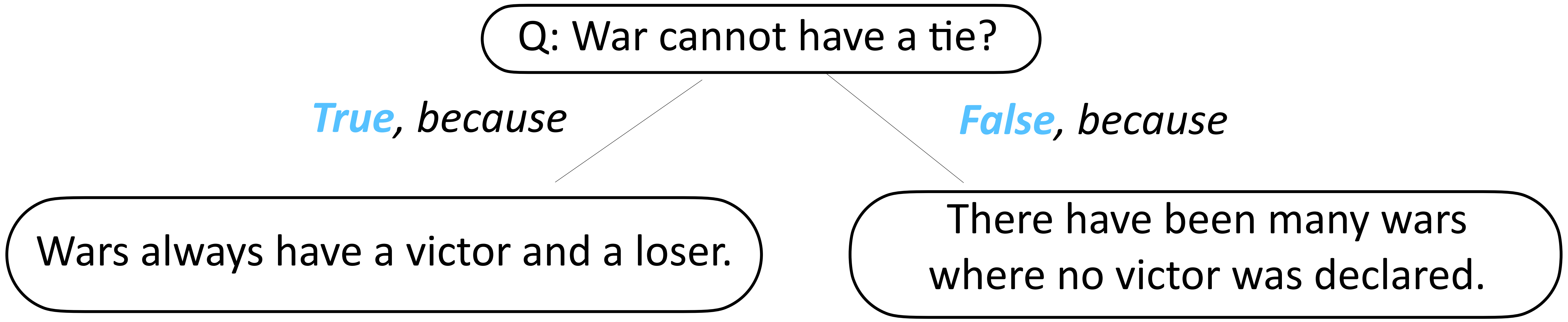
Problem setting

- Binary labels
    - x: text
    - $y \in \{0,1\}$

- True/False question answering
- Claim verification

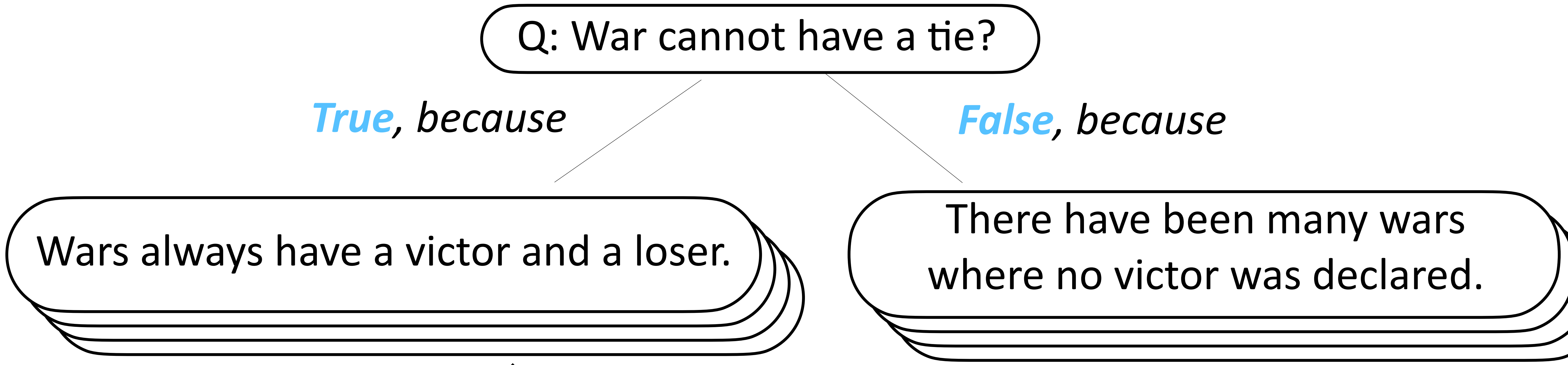
Method | enumerate tree

- Label-conditioned generation
- $e_{1,a} \sim p(e \mid a, q; D)$



Method | enumerate tree

- Label-conditioned generation
- $e_{1,a} \sim p(e \mid a, q; D)$

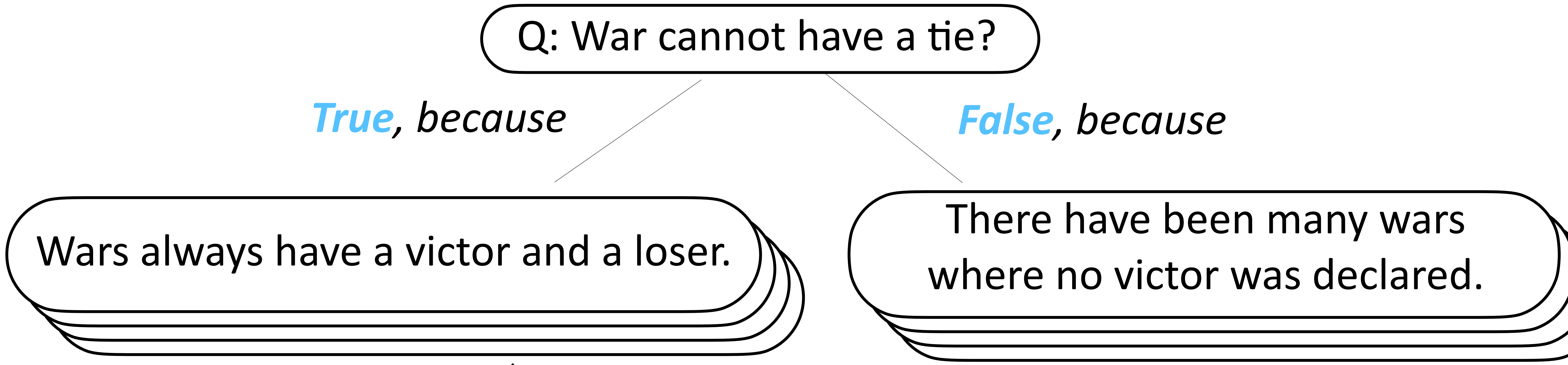


- Multiple samples

Method | enumerate tree

- Label-conditioned generation
- $e_{1,a} \sim p(e \,|\, a, q; D)$
  - Prompt (6 training examples)
    -

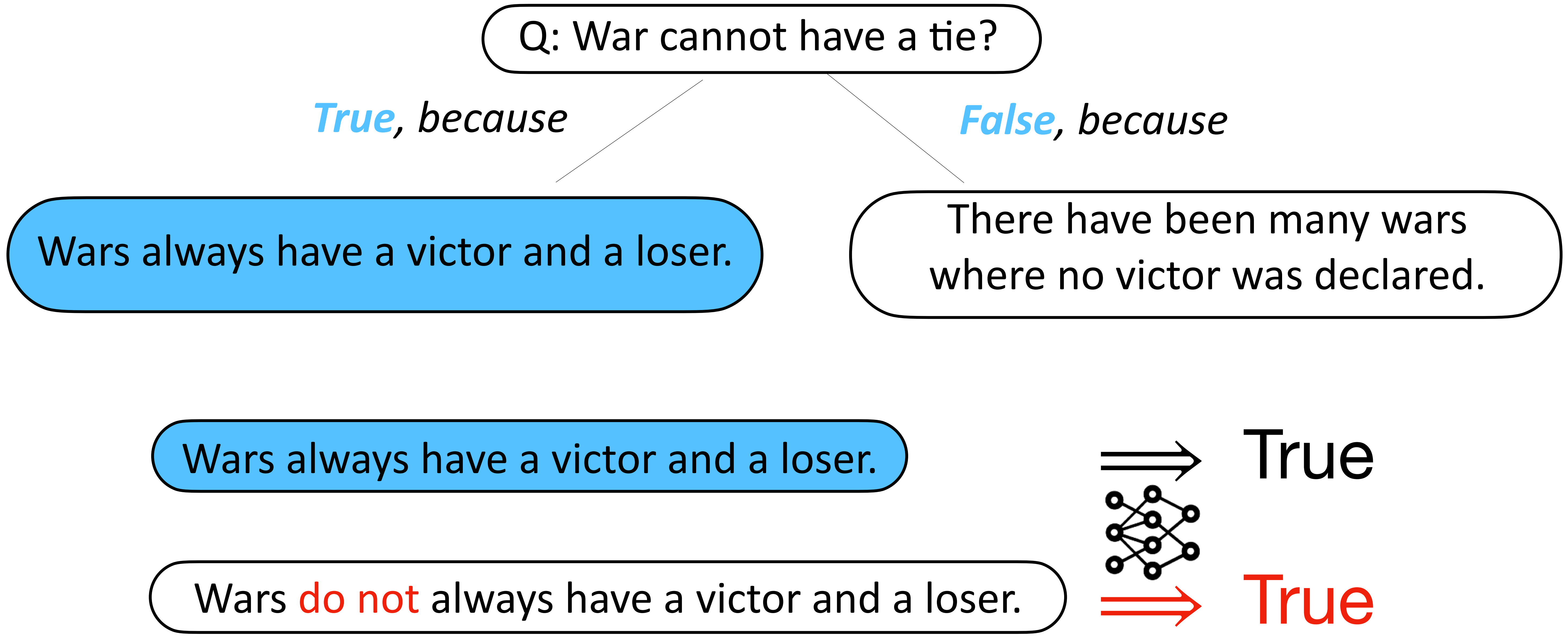
```
Given a statement, determine whether the statement makes sense, and explain the reason.
###
Q: Jane loved to upset her parents with loud noises so she threw a paper plate on the floor?
A: This statement is false, because a paper plate is light and will not make any noise when thrown to the floor.
###
Q: To see stars at night, it is better not to turn on the lights?
A: This statement is true, because Stars are seen more clearly when it's dark.
###
Q: If you want a drink that wakes you up, it's better to look for one with a lot of caffeine rather than protein?
A: This statement is true, because caffeine is a stimulant and will wake you up.
###
Q: It was January in New York so Pat knew that he would see more people at the park rather than in the gym?
A: This statement is false, because it's usually freezing in New York on January, so people would prefer staying indoor rather than going outside.
###
Q: A man who can bench press two hundred pounds can easily lift a small child?
A: This statement is true, because a small child typically weighs way less than 200 pounds.
###
Q: It is a hot day, so Fenton grabbed a big, red popsicle. If Fenton doesn't want to stain the floor, he should stand in the room with the carpeted floor?
A: This statement is false, because if one spills popsicle to the carpet, it will be difficult to clean up because the carpet will absorb it.
###
```



- Multiple samples

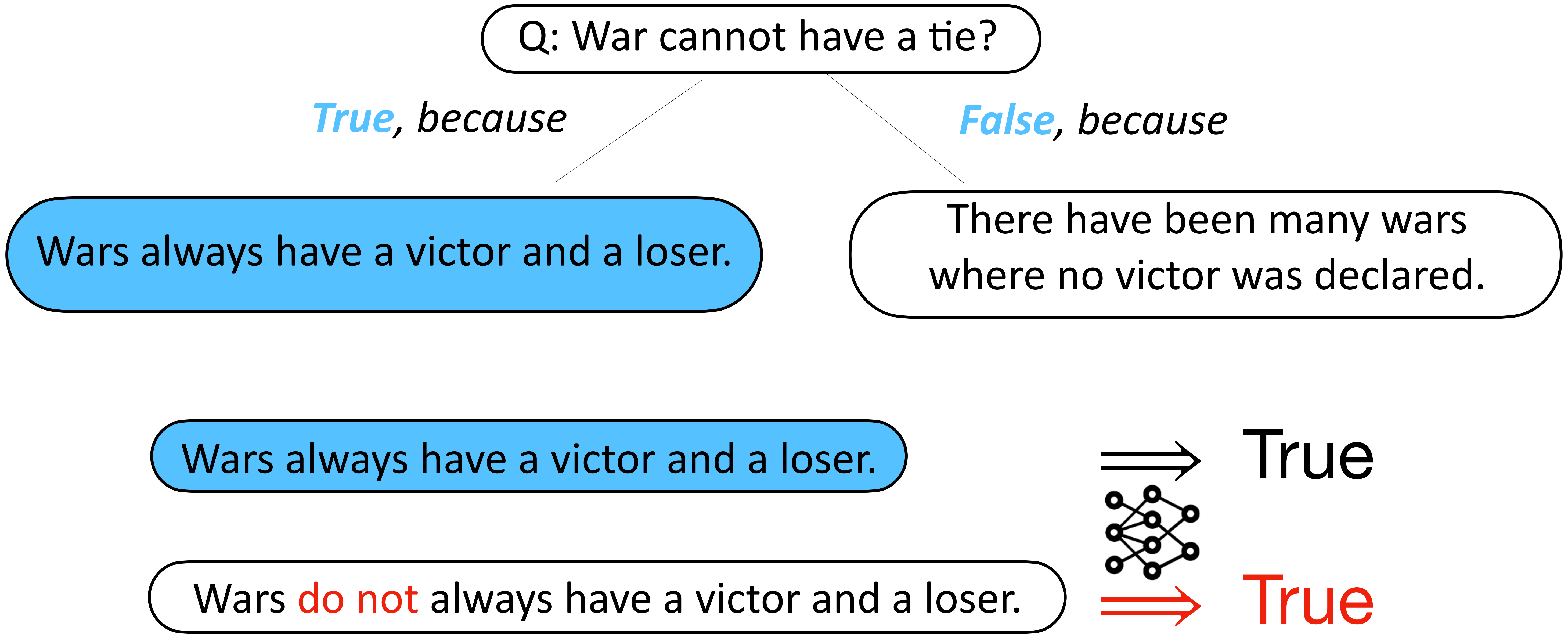
Method | enumerate tree

- Check logical integrity of claim



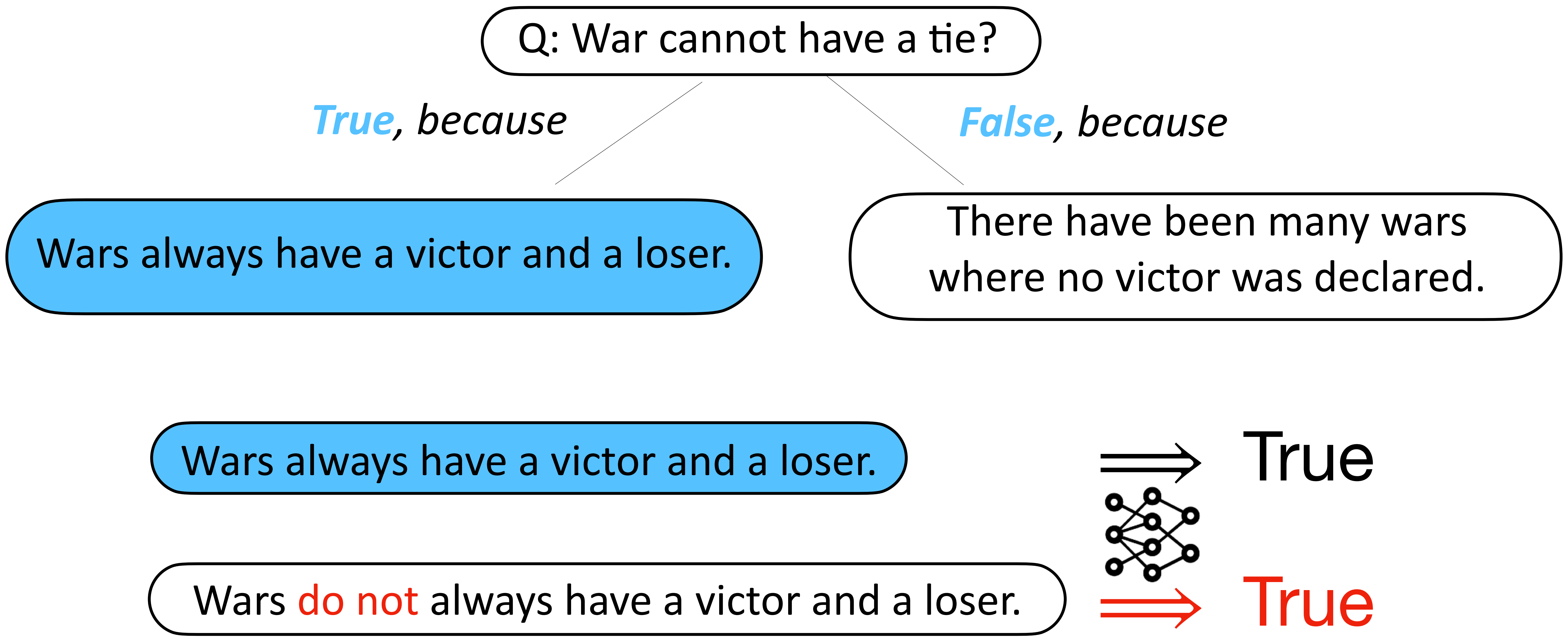
Method | enumerate tree

- Check logical integrity of claim
- Does the LM predict
  True given E, **False** given $\neg E$



Method | enumerate tree

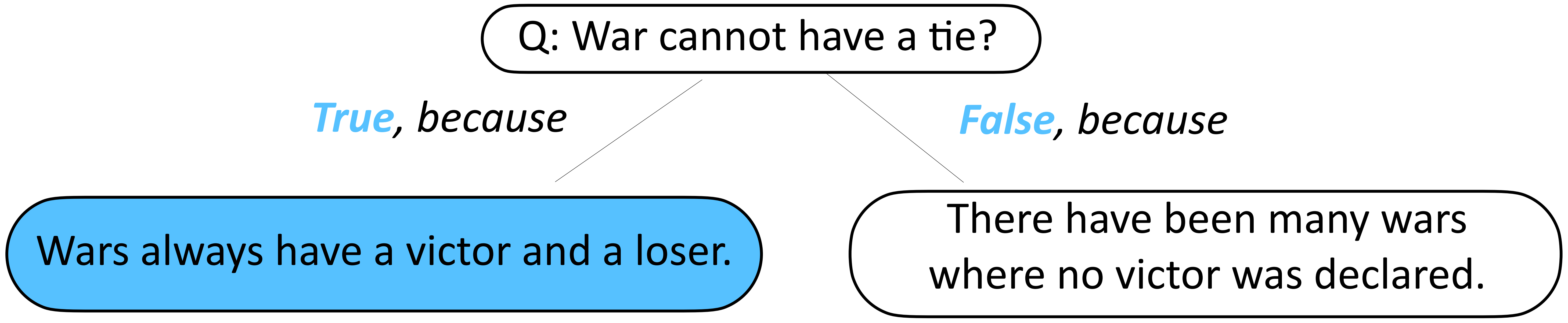
- Check logical integrity of claim
- Does the LM predict **True** given E, **False** given $\neg E$
- $p(\{T, F\} \mid e; D)$

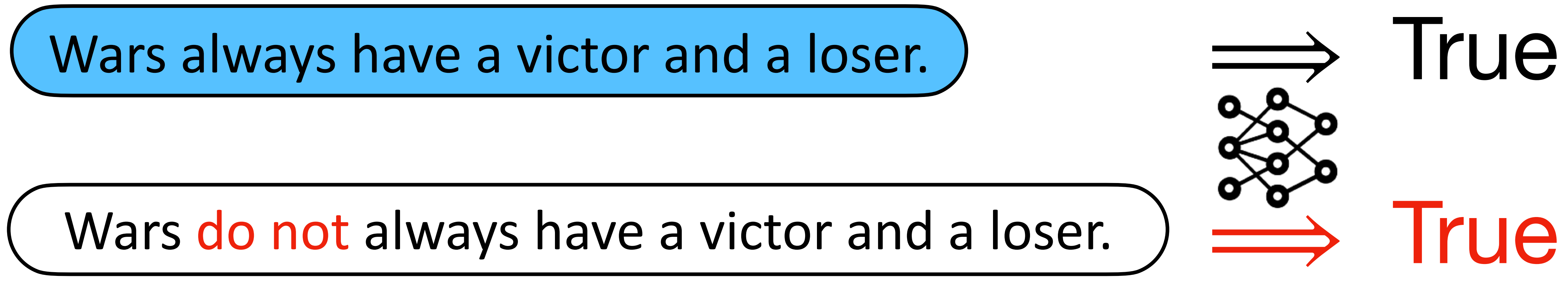


Method | enumerate tree

- Check logical integrity of claim
- Does the LM predict
  True given E, **False** given $\neg E$

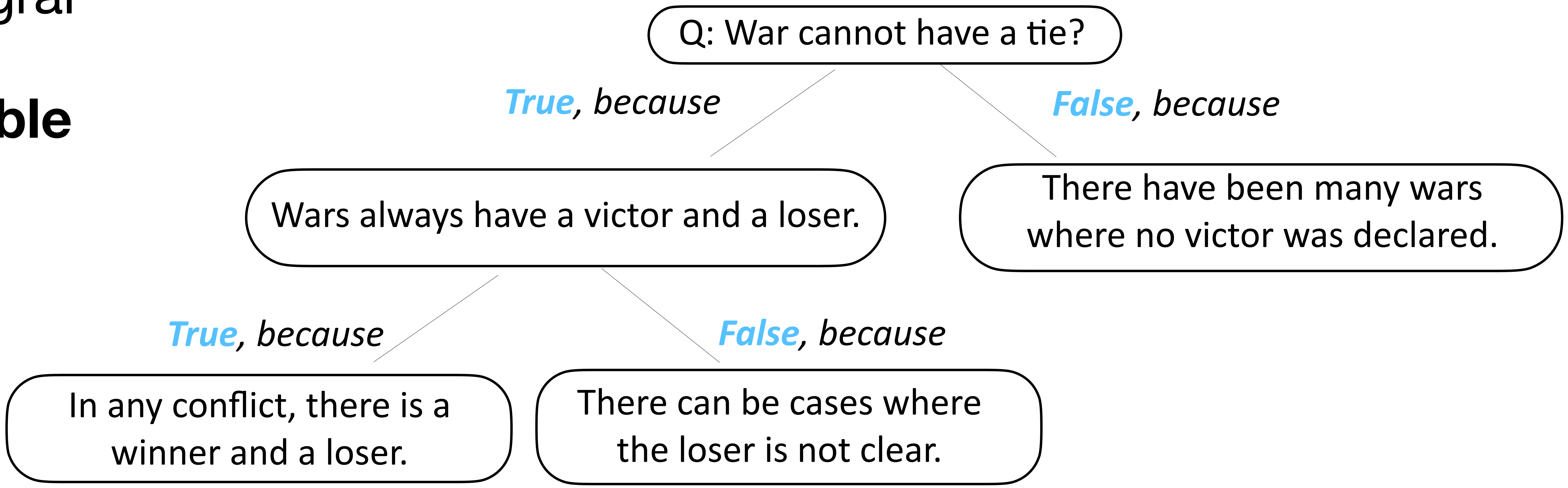
- $p(\{T, F\} \mid e; D)$
- $p(\{T, F\} \mid \neg e; D)$
    - Again, just prompts





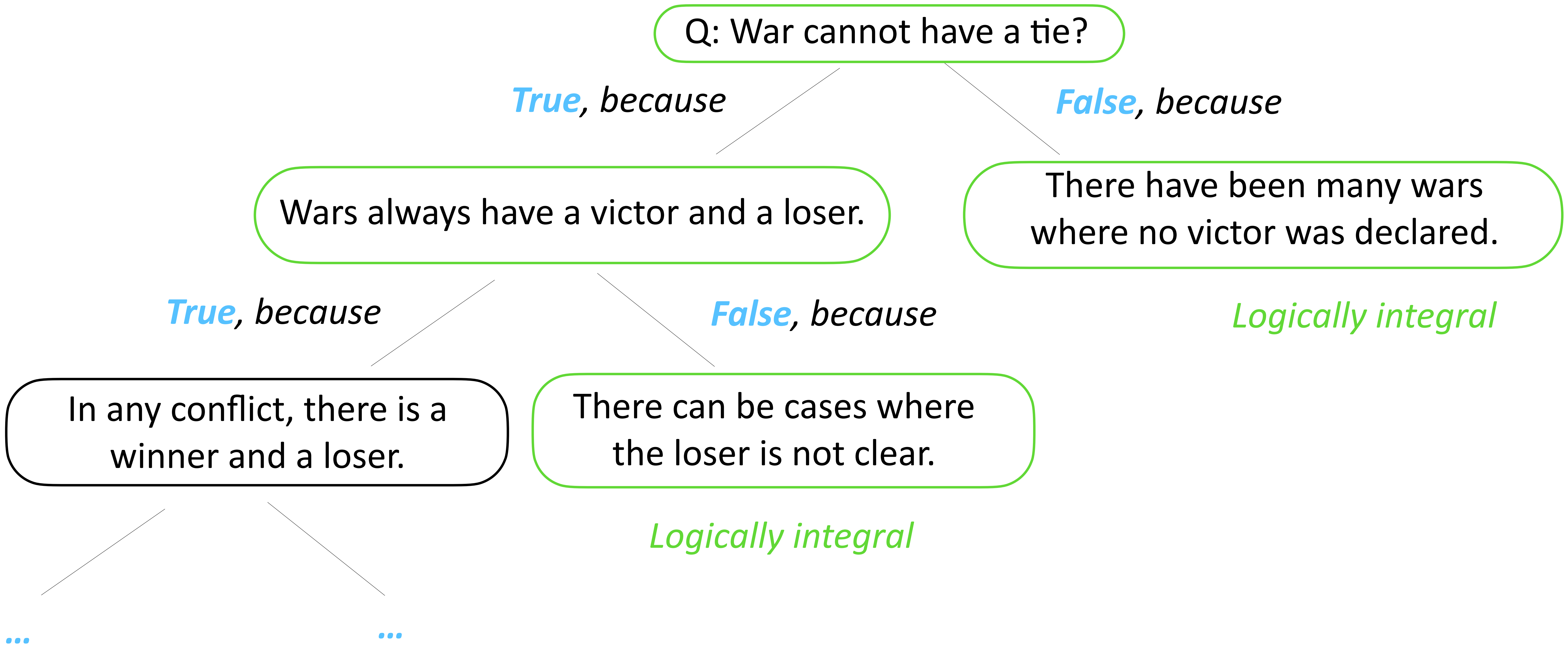
Method | enumerate tree

- Expand if not logically integral
  - $p(\{T, F\} \mid e)$ is **not reliable**



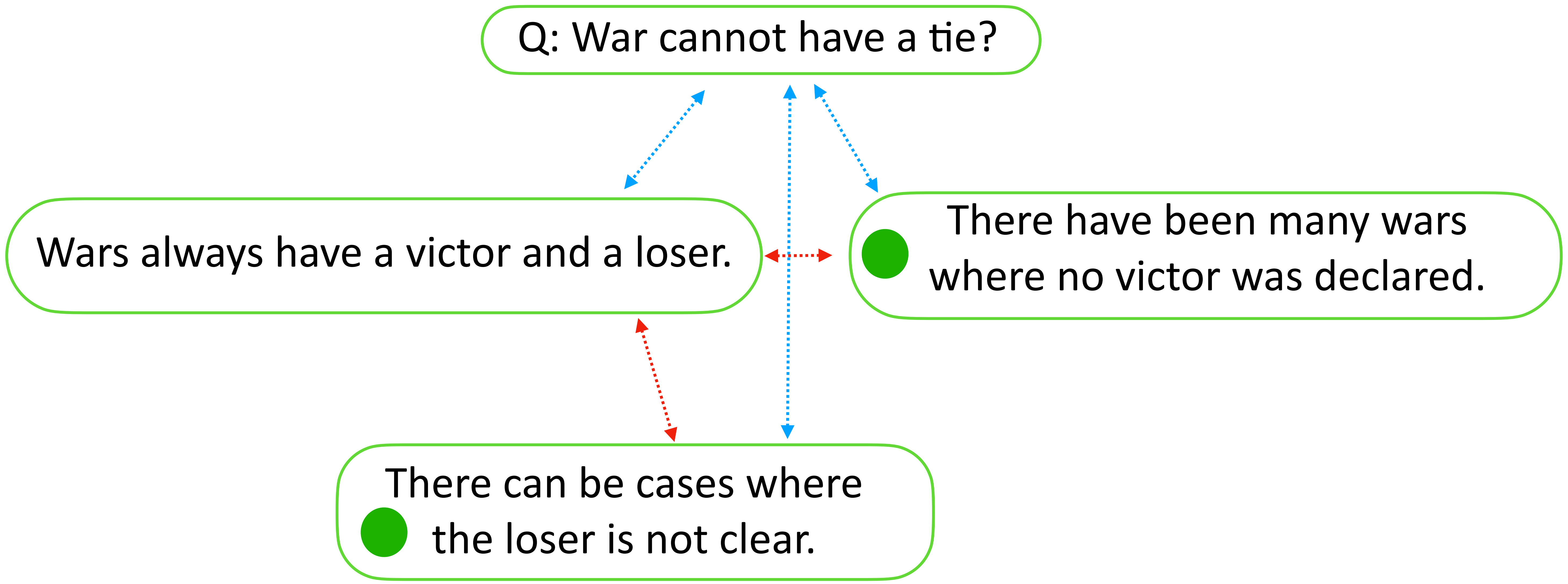
Method | enumerate tree

- Stop if logically integral
  - $p(\{T, F\} \mid e)$ is **reliable**



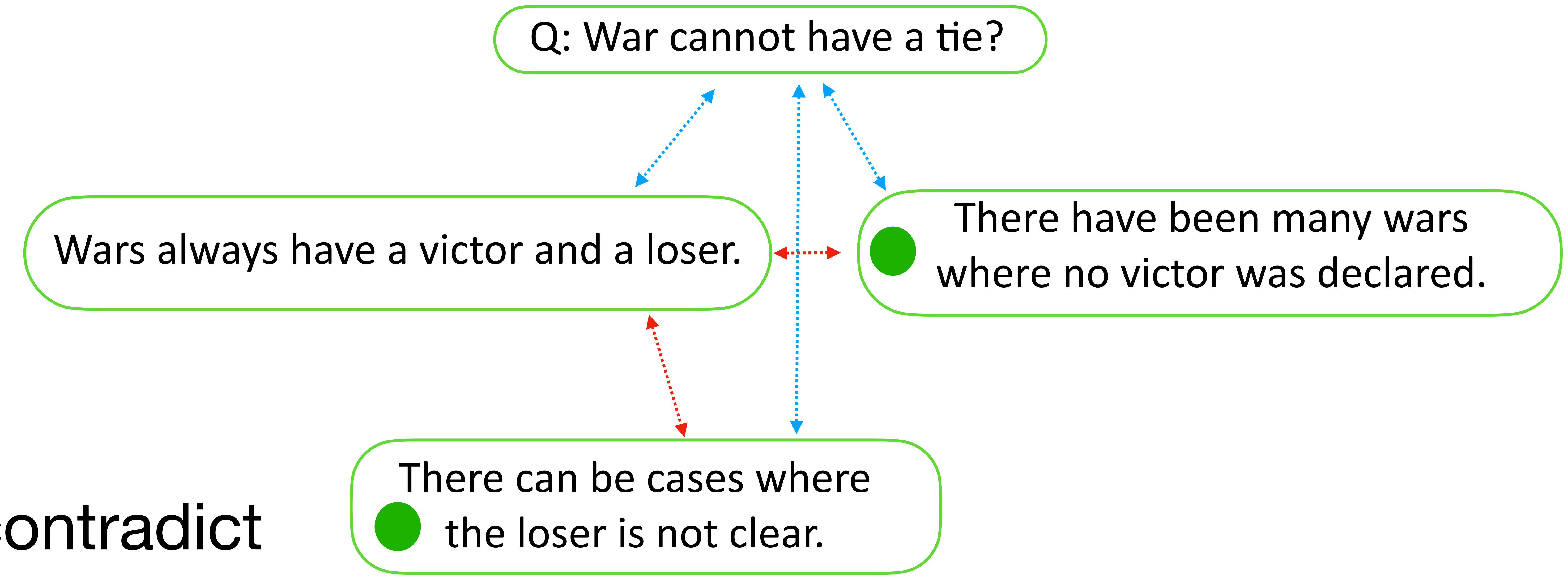
Method | scoring

- Logically integral nodes:
  - $w_e = p(T \mid e; D)$
  - "Model's belief about claim"



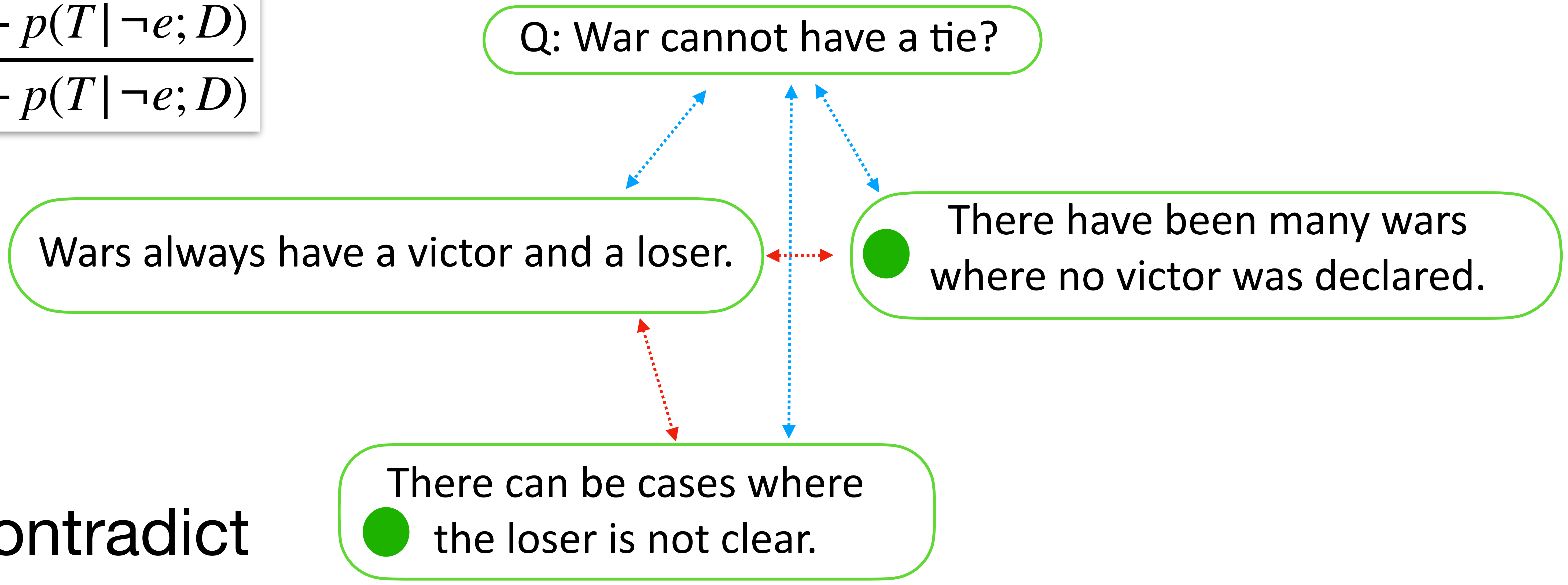
Method | scoring

- Logically integral nodes:
  - $w_e = p(T \mid e; D)$
  - "Model's belief about claim"
- Relations:
  - $w_{e_i,e_j}: f(e_i, e_j) \rightarrow$ entail, neutral, contradict
  - Off-the-shelf NLI model
  - "Internal contradictions"



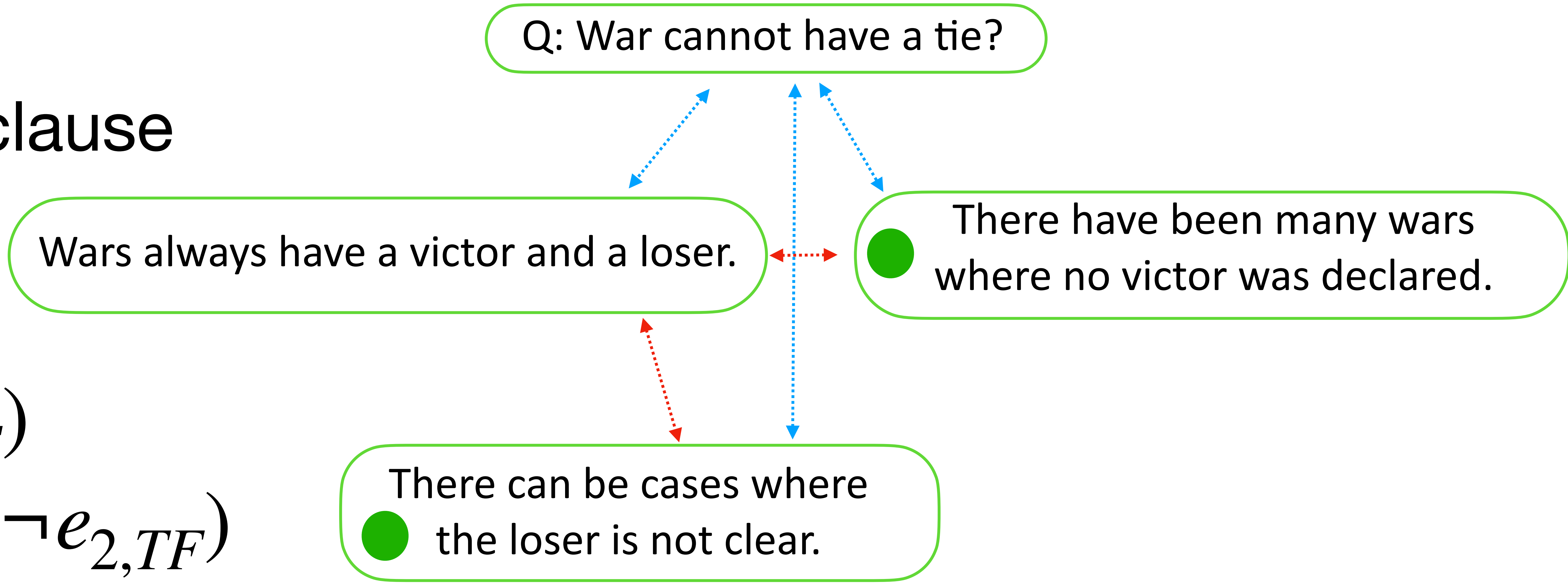
Method | scoring

- Logically integral nodes:
  - $w_e = p(T \mid e; D)$ $\quad \dfrac{p(T \mid e; D) - p(T \mid \neg e; D)}{p(T \mid e; D) + p(T \mid \neg e; D)}$
  - “Model’s belief about claim”
- Relations:
  - w_{e_i,e_j}: $f(e_i, e_j) \rightarrow$ entail, neutral, contradict
  - Off-the-shelf NLI model
  - “Internal contradictions”



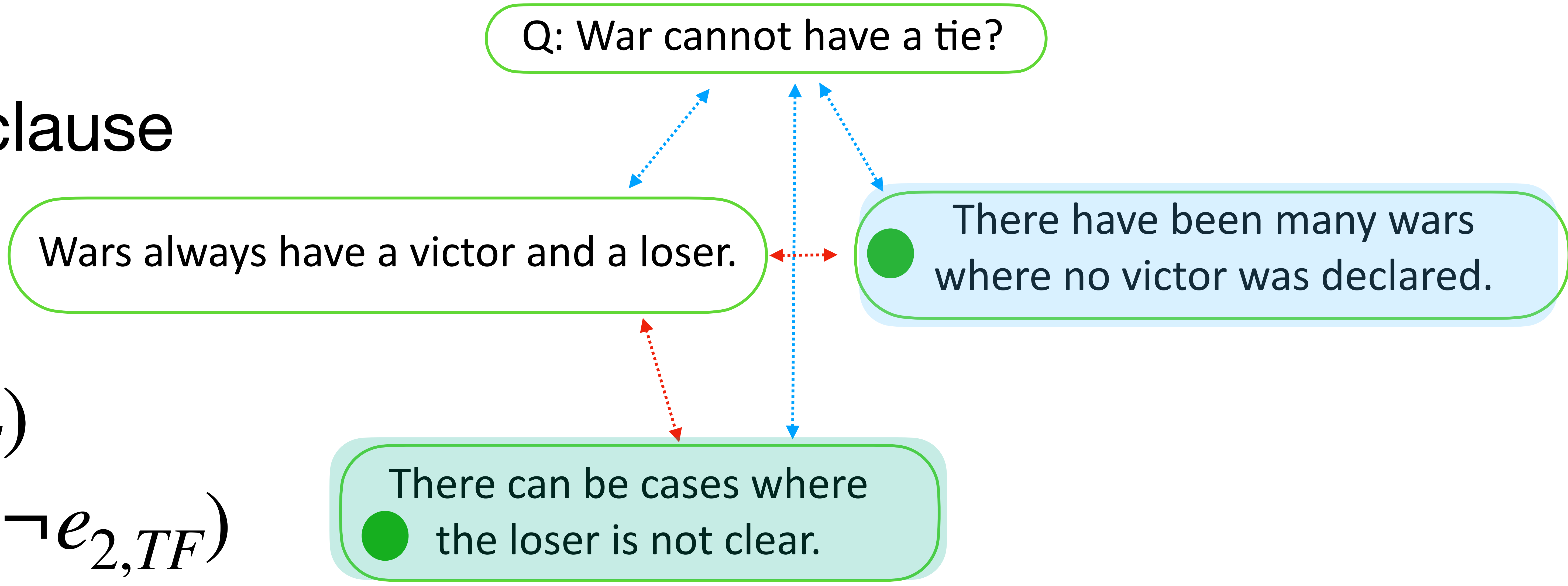
Method | aggregation

- Tree: **weighted CNF formula**
  - **Logically integral node:** unary clause
  - **NLI:** implication clause
- $w_{1,F} \cdot (e_{1,T}) \land w_{q1F} \cdot (q \implies e_{1,F})$
  $\land\, w_{2,TF} \cdot (e_{2,TF}) \land w_{...}(e_{2,T} \implies \neg e_{2,TF})$
  $\ldots$



Method | aggregation

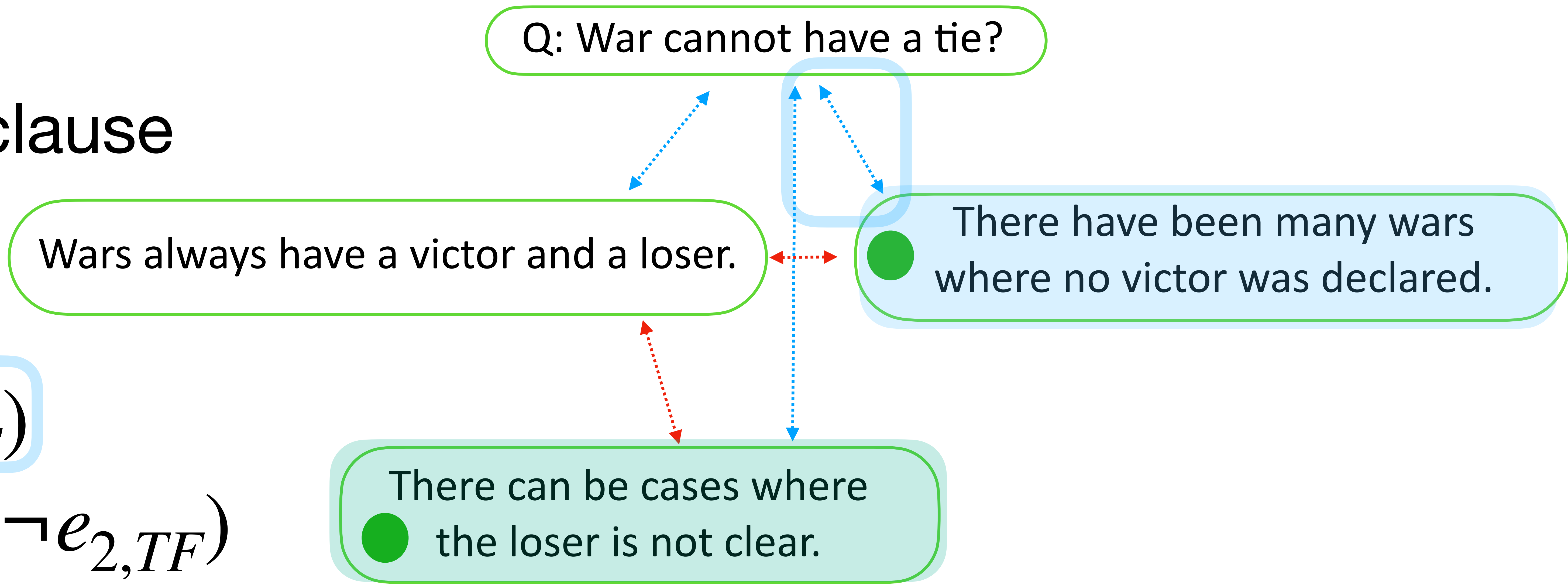
- Tree: **weighted CNF formula**
  - **Logically integral node:** unary clause
  - **NLI:** implication clause



- $w_{1,F} \cdot (e_{1,T}) \land w_{q1F} \cdot (q \implies e_{1,F})$
  $\land\, w_{2,TF} \cdot (e_{2,TF}) \land w_{...}(e_{2,T} \implies \neg e_{2,TF})$
  $\ldots$

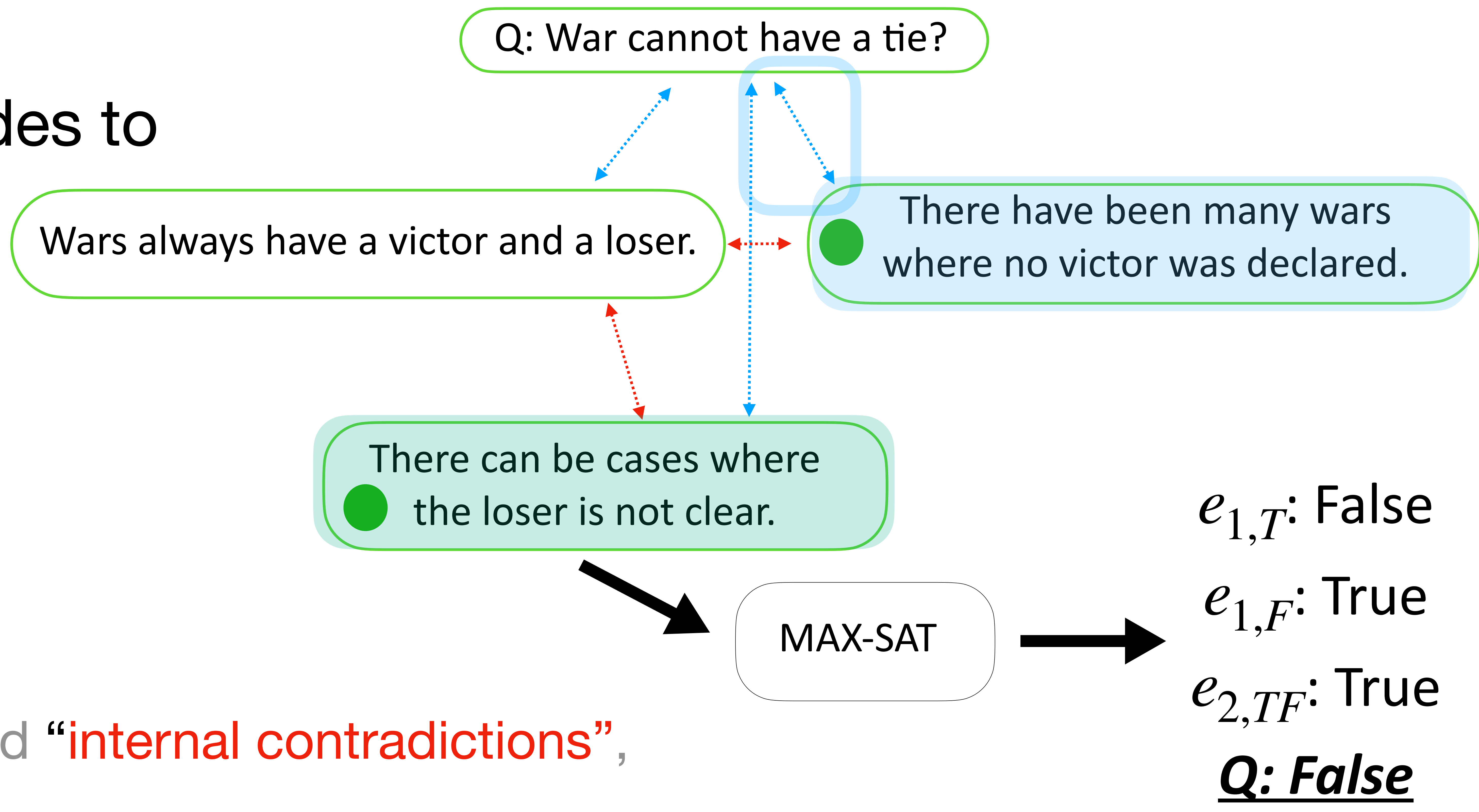
Method | aggregation

- Tree: **weighted CNF formula**
  - **Logically integral node:** unary clause
  - **NLI:** implication clause
- $w_{1,F} \cdot (e_{1,T}) \wedge w_{q1F} \cdot \boxed{(q \implies e_{1,F})}$
  $\wedge\, w_{2,TF} \cdot (e_{2,TF}) \wedge w_{...}(e_{2,T} \implies \neg e_{2,TF})$
  ...



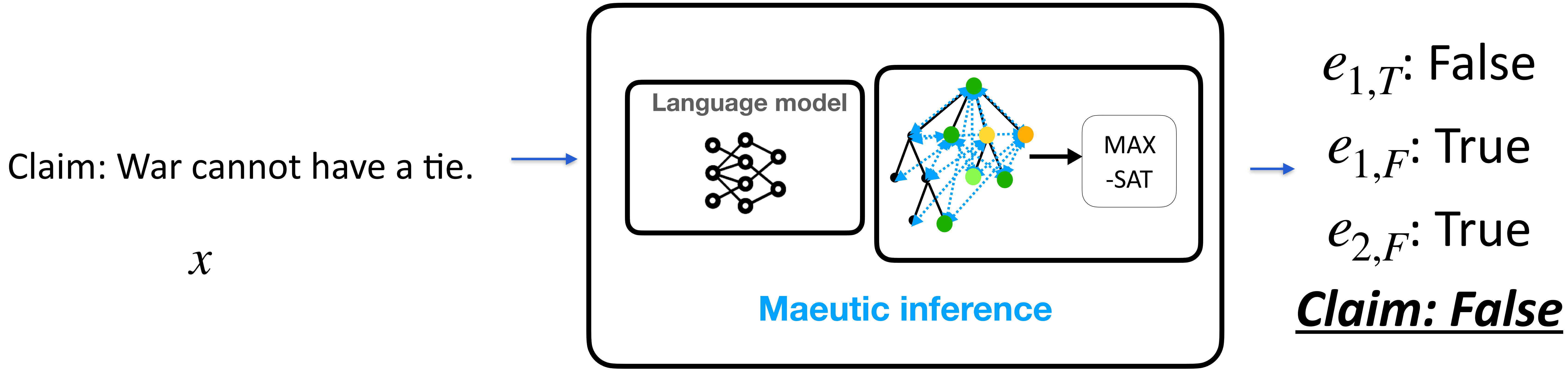
Method | aggregation

- Tree: **weighted CNF formula**
- **MAX-SAT**: Assign true/false to nodes to maximize total weight



- Intuition: Resolve "belief about claims" and "internal contradictions", into a decision about which ones are true

Method | Maeutic inference



1. Enumerate tree of explanations
2. Score relations in tree
3. Resolve scores into a prediction

Experiments

Experiments

- Commonsense reasoning / fact verification:
  - Com2Sense
    Commonsense QA 2.0
    CREAK

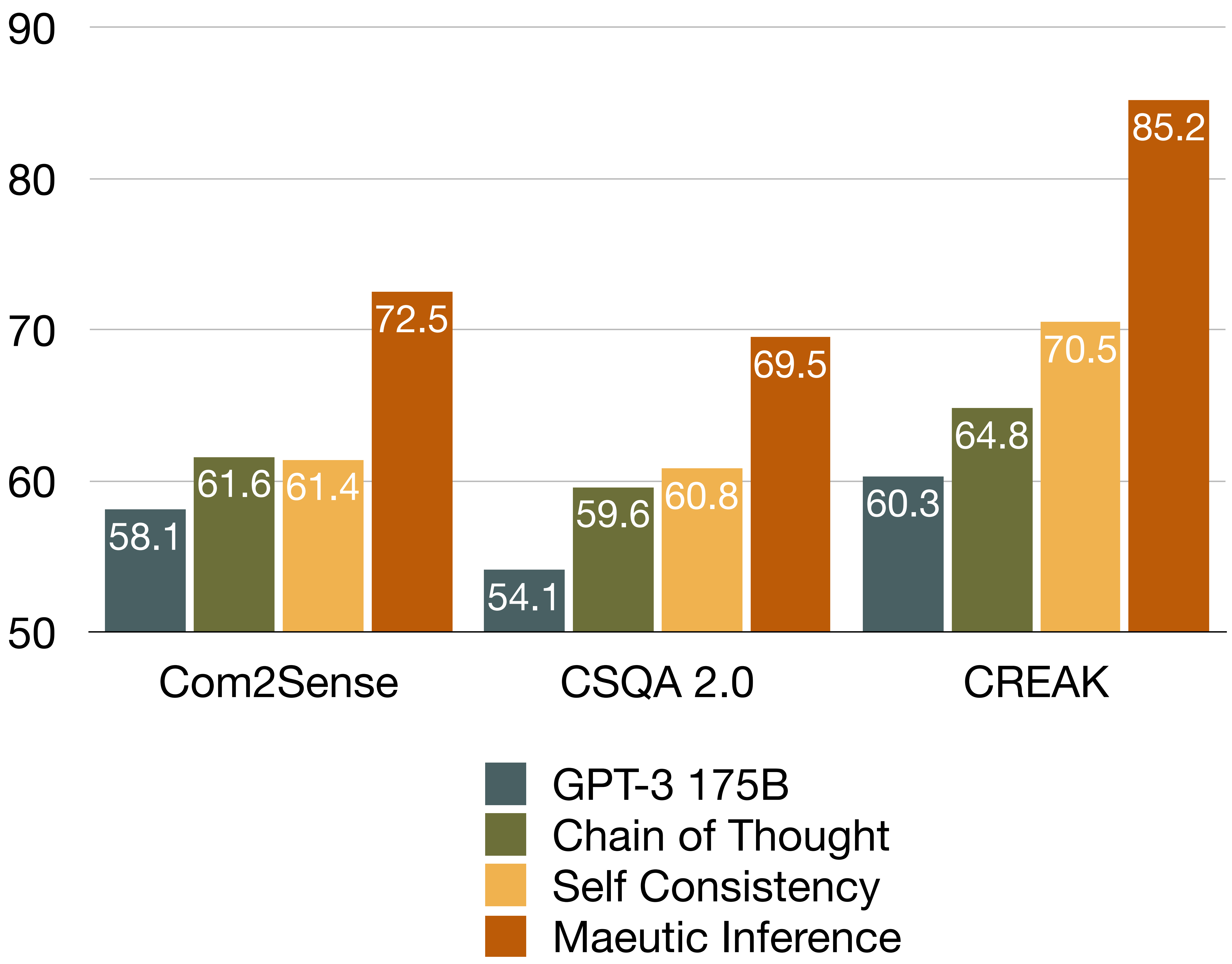
Experiments

- Commonsense reasoning / fact verification:
  - Com2Sense
    Commonsense QA 2.0
    CREAK
- Model:
  - GPT3 (text-davinci-001), with 6-shot prompt per dataset
  - NLI Model: Roberta fine-tuned on MNLI

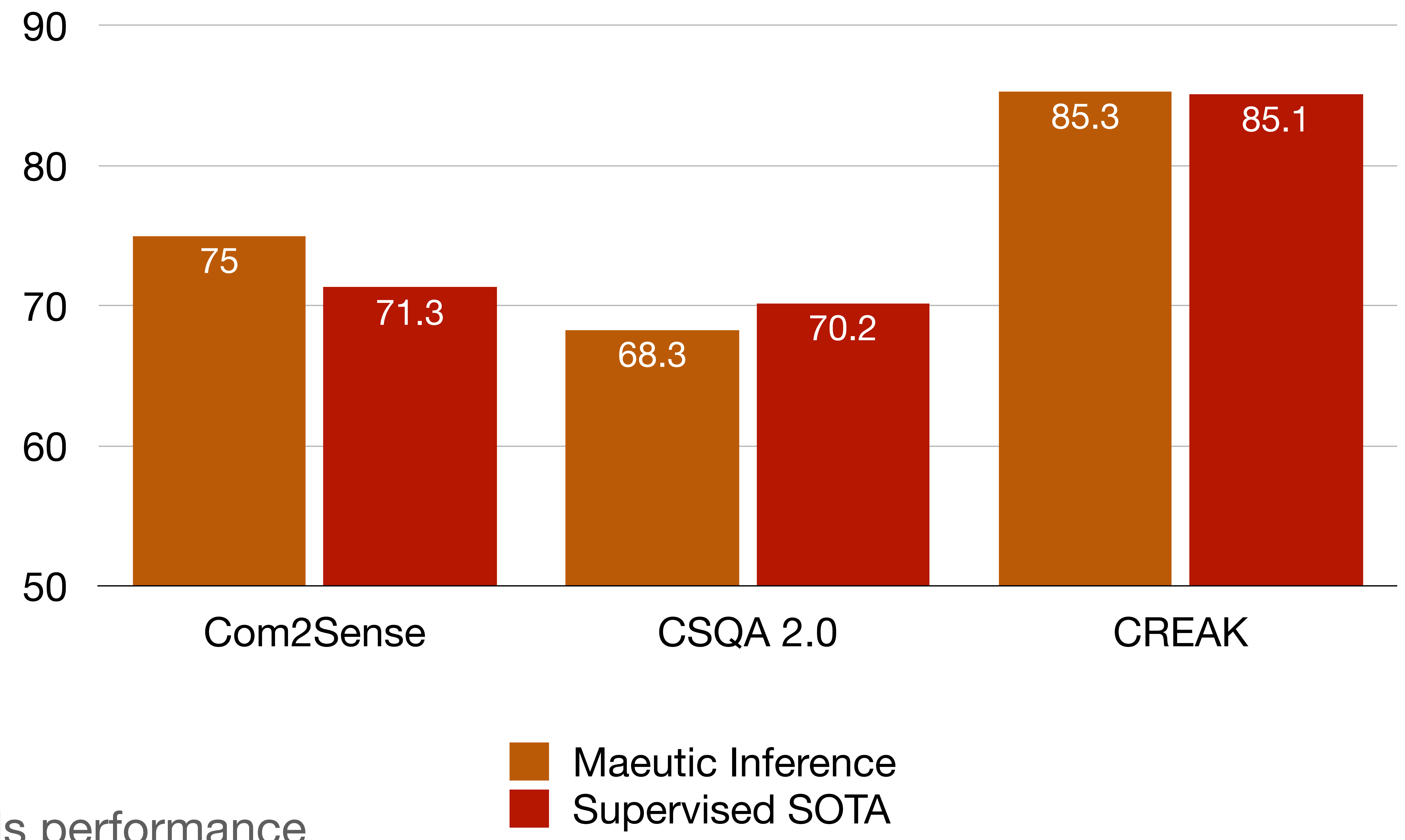
Experiments

- Commonsense reasoning / fact verification:
  - Com2Sense
    Commonsense QA 2.0
    CREAK
- Model:
  - GPT3 (text-davinci-001), with 6-shot prompt per dataset
  - NLI Model: Roberta fine-tuned on MNLI
- Settings:
  - 3 True/3 False expansions, then 1 greedy recursive expansion (max 18 nodes)

Benchmark performance

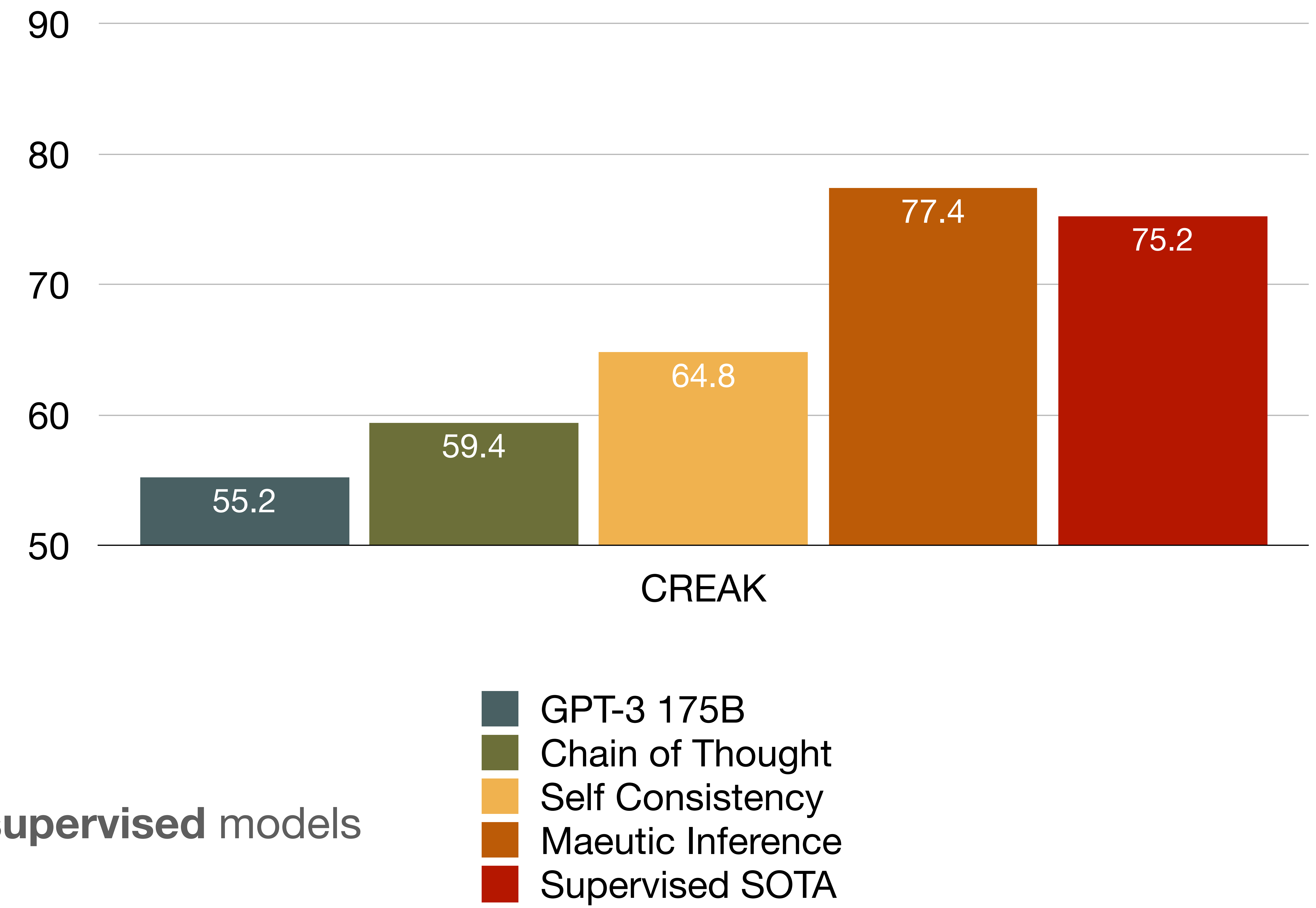


Benchmark performance



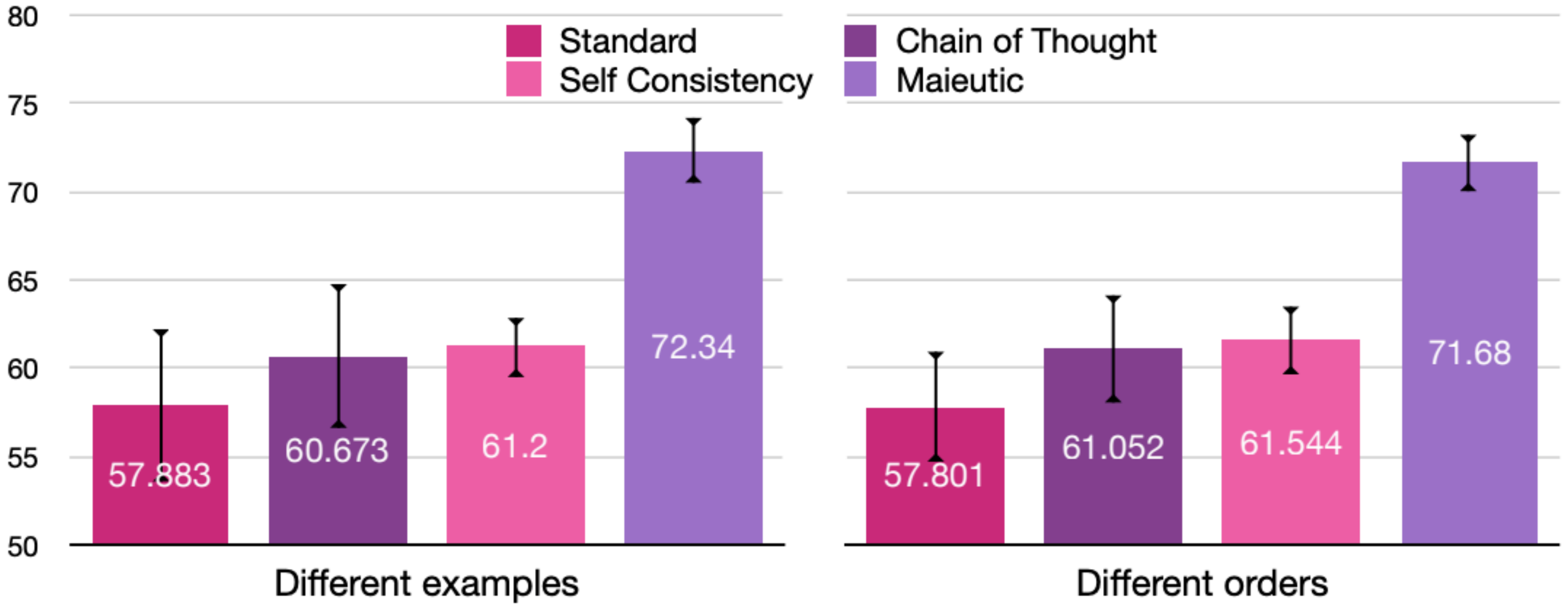
~ approaches/exceeds performance
of **supervised** models!

Robustness



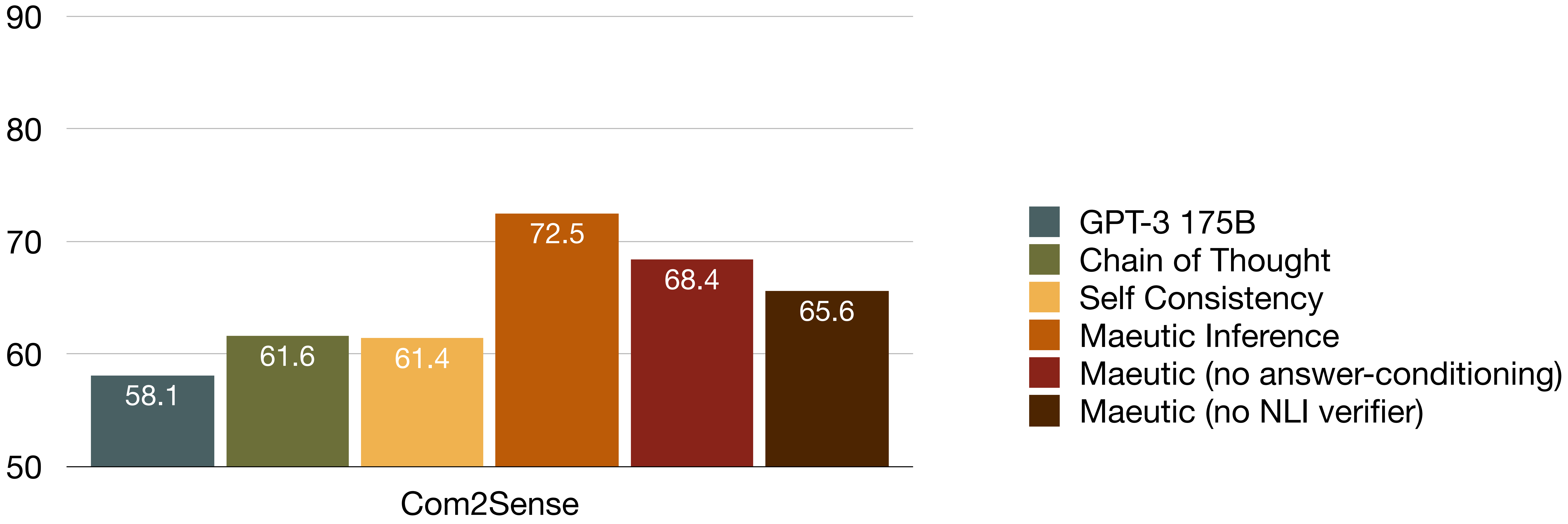
+ more robust than **supervised** models

Robustness



80
75
70
65
60
55
50
Standard
Self Consistency
Chain of Thought
Maieutic
57.883
60.673
61.2
72.34
57.801
61.052
61.544
71.68
Different examples
Different orders

Ablations



Answer conditioning & verifier important
(but still beats baselines without)

Ablations

Dimension	1	2	3	5	10
Depth	61.3	72.5	72.4	-	-
Width	62.4	66.5	72.5	71.5	72.1

Table 3: Performance of MAIEUTIC PROMPTING on Com2Sense with different maieutic tree sizes.

Interpretability

Q: If you travel west far enough from the west coast, you will reach the east coast.

E_{T_0}: The Earth is round and if you travel in any direction long enough, you will eventually return to where you started.
integral / True

E_{T_2}: All directions eventually meet at the North and South Poles.
integral / True

E_{F_0}: You cannot reach the east coast by going west.

E_{F_2}: You can only travel so far before you reach the end of the earth.

E_{T_1}: The world is round and if you continue to travel in a straight line, you will eventually reach the other side.

$E_{F_0T_0}$: You can reach the east coast by going west by traveling around the world.
integral / False

$E_{F_2F_0}$: The Earth is round.
integral / True

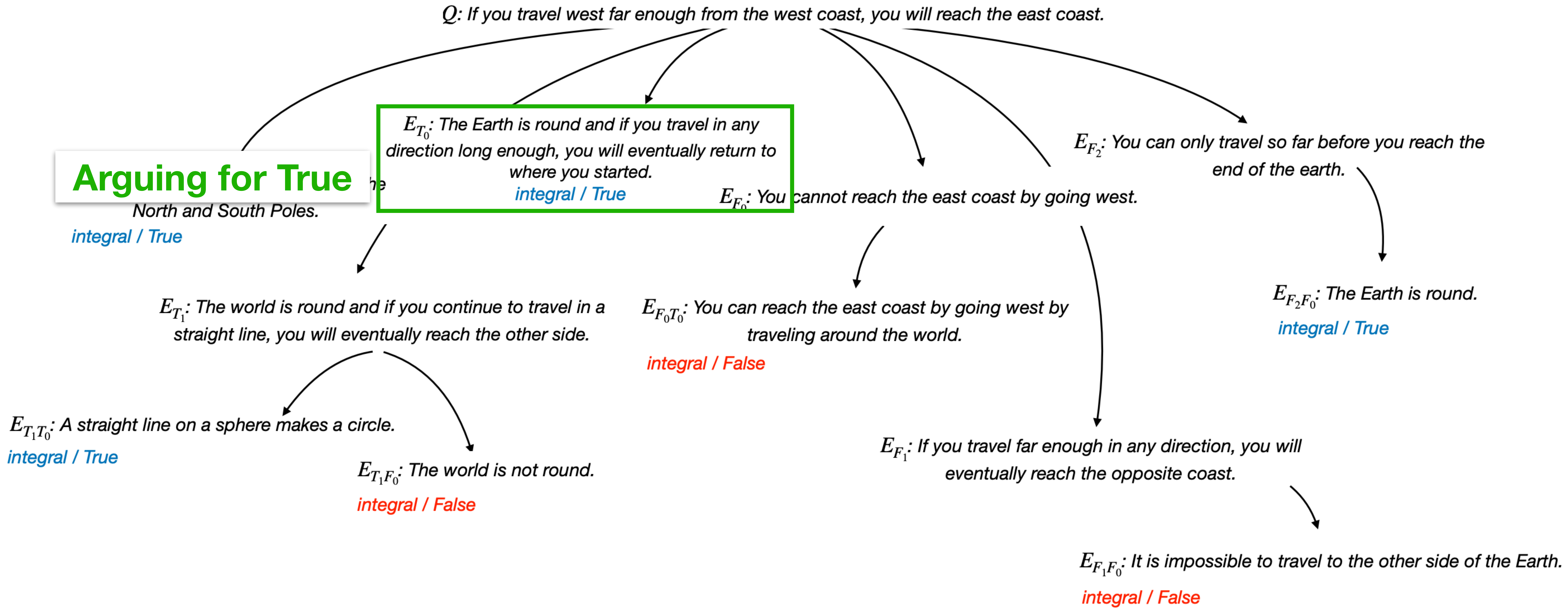
$E_{T_1T_0}$: A straight line on a sphere makes a circle.
integral / True

$E_{T_1F_0}$: The world is not round.
integral / False

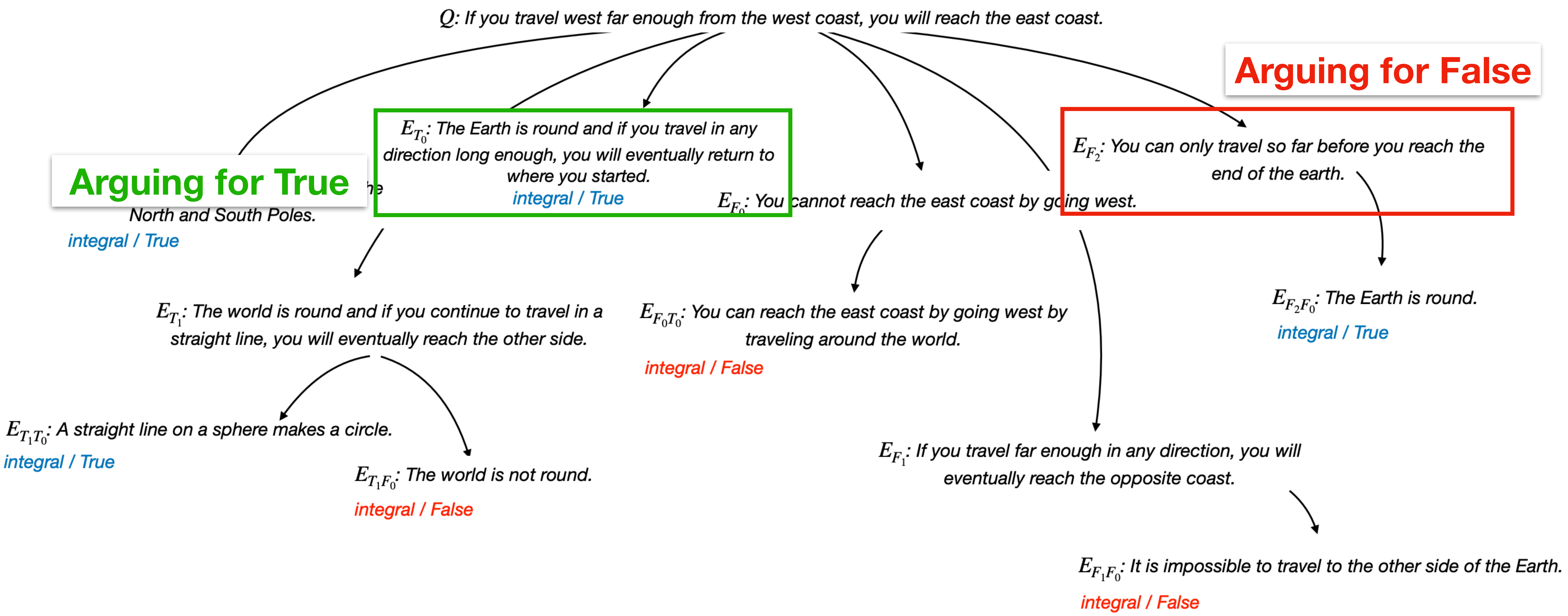
E_{F_1}: If you travel far enough in any direction, you will eventually reach the opposite coast.

$E_{F_1F_0}$: It is impossible to travel to the other side of the Earth.
integral / False

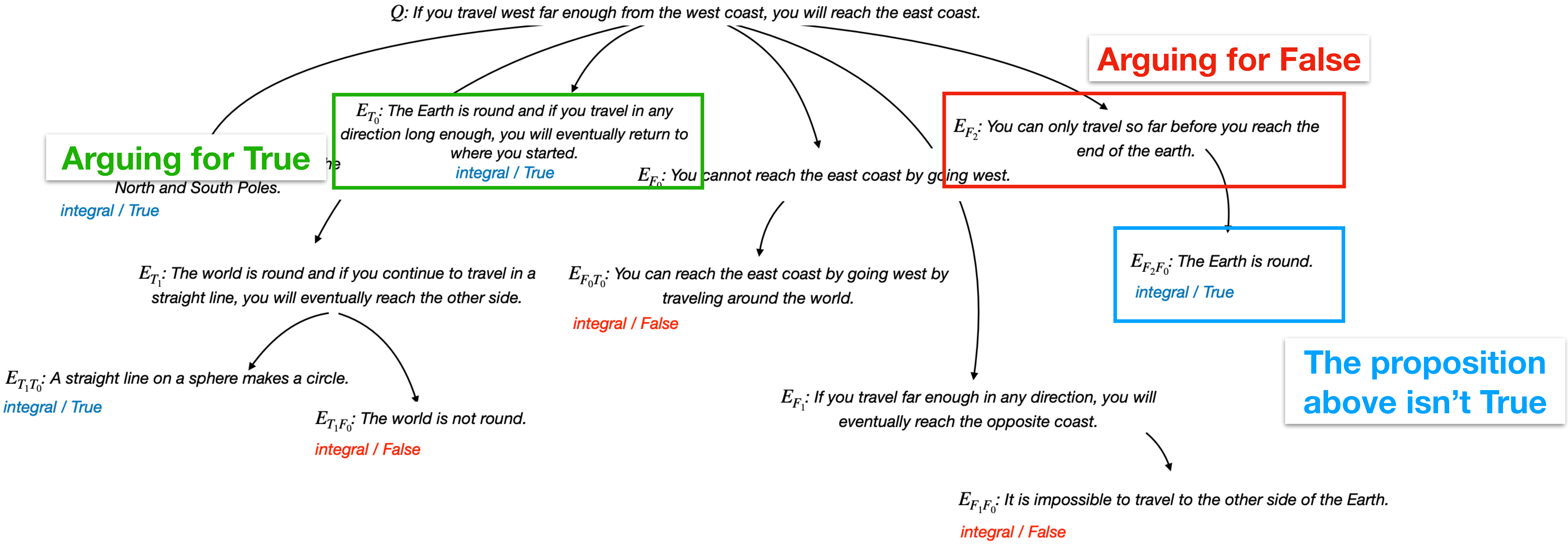
Interpretability



Interpretability

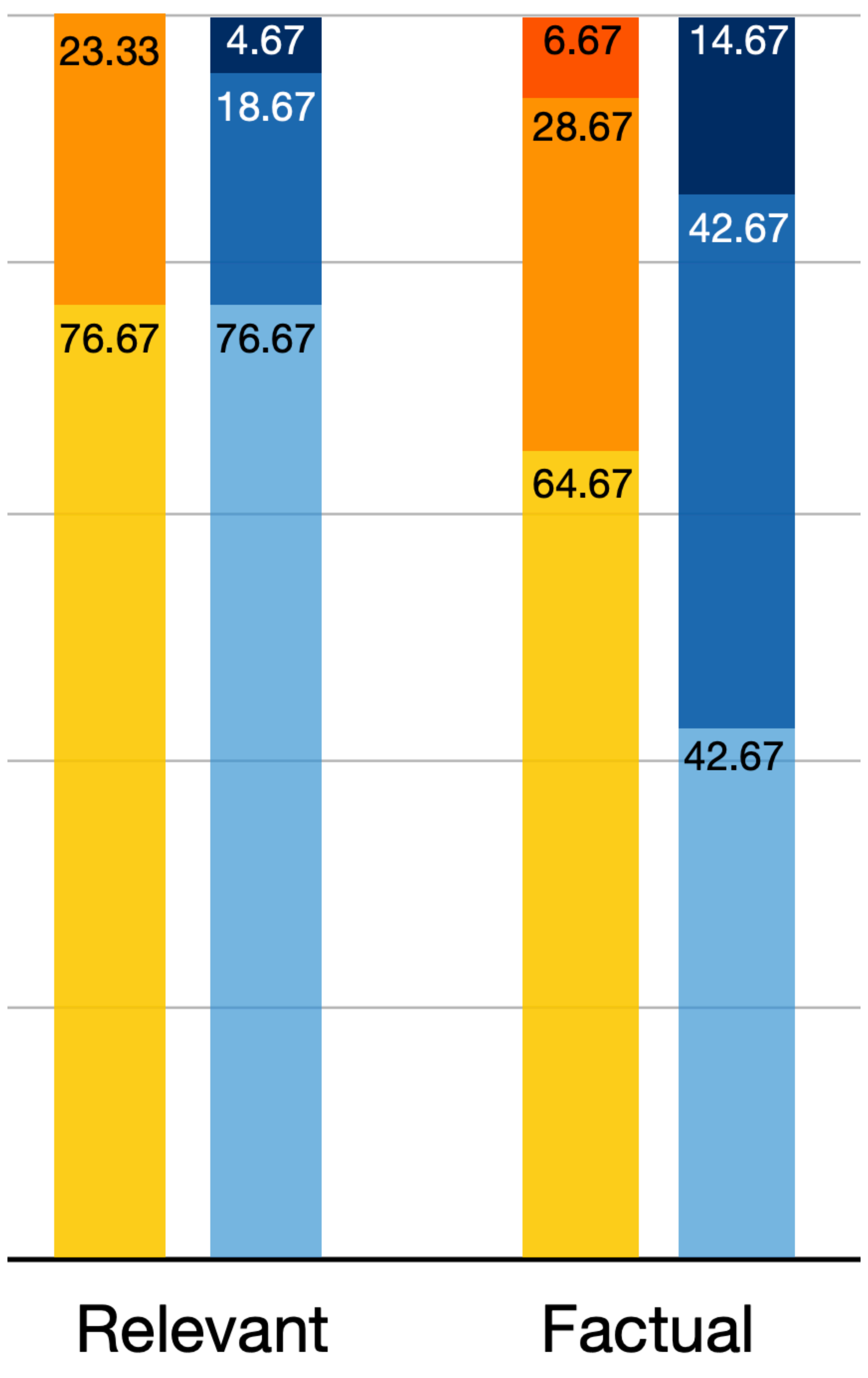


Interpretability



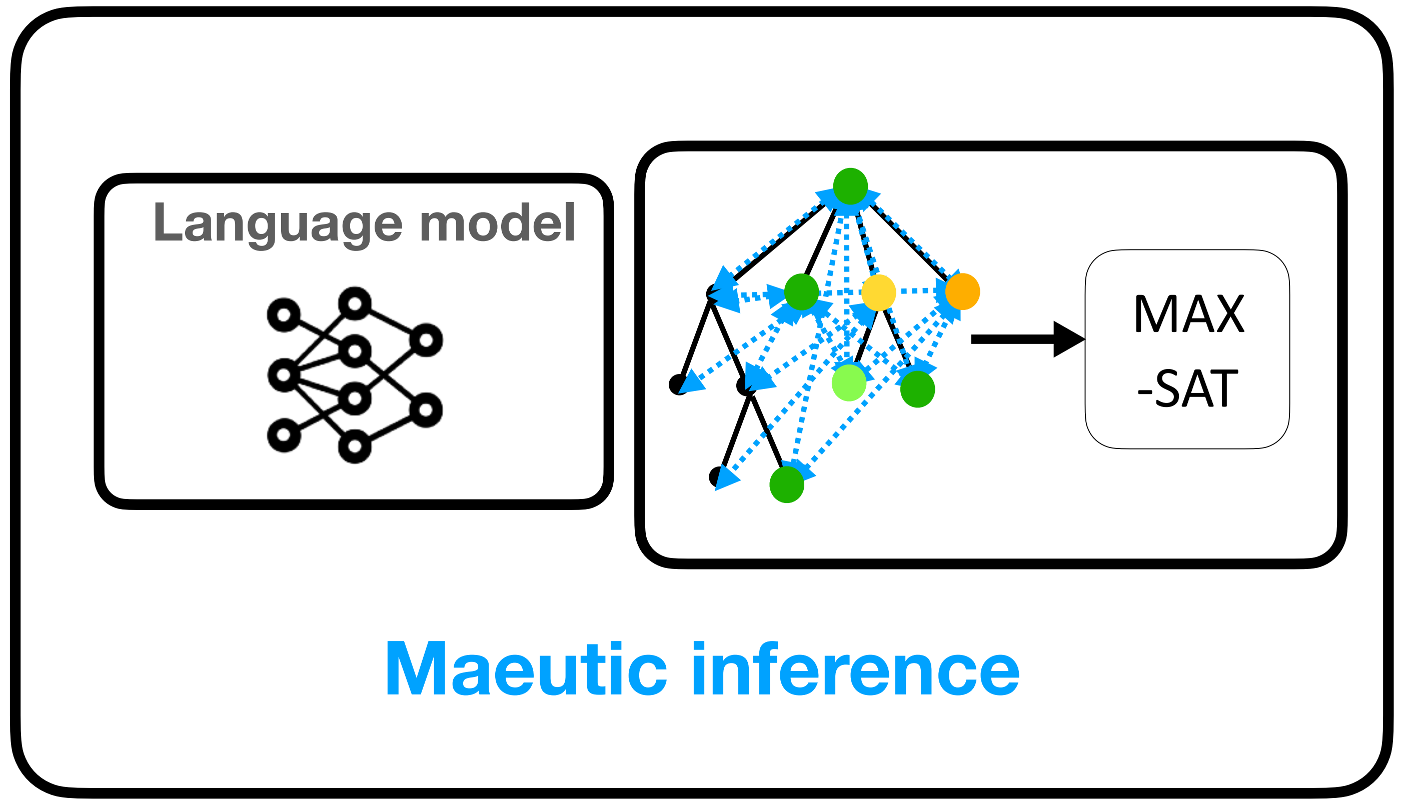
Interpretability

- Propositions identified by MAX-SAT are typically relevant and factual
  - Even when the answer is incorrect! **(Blue)**



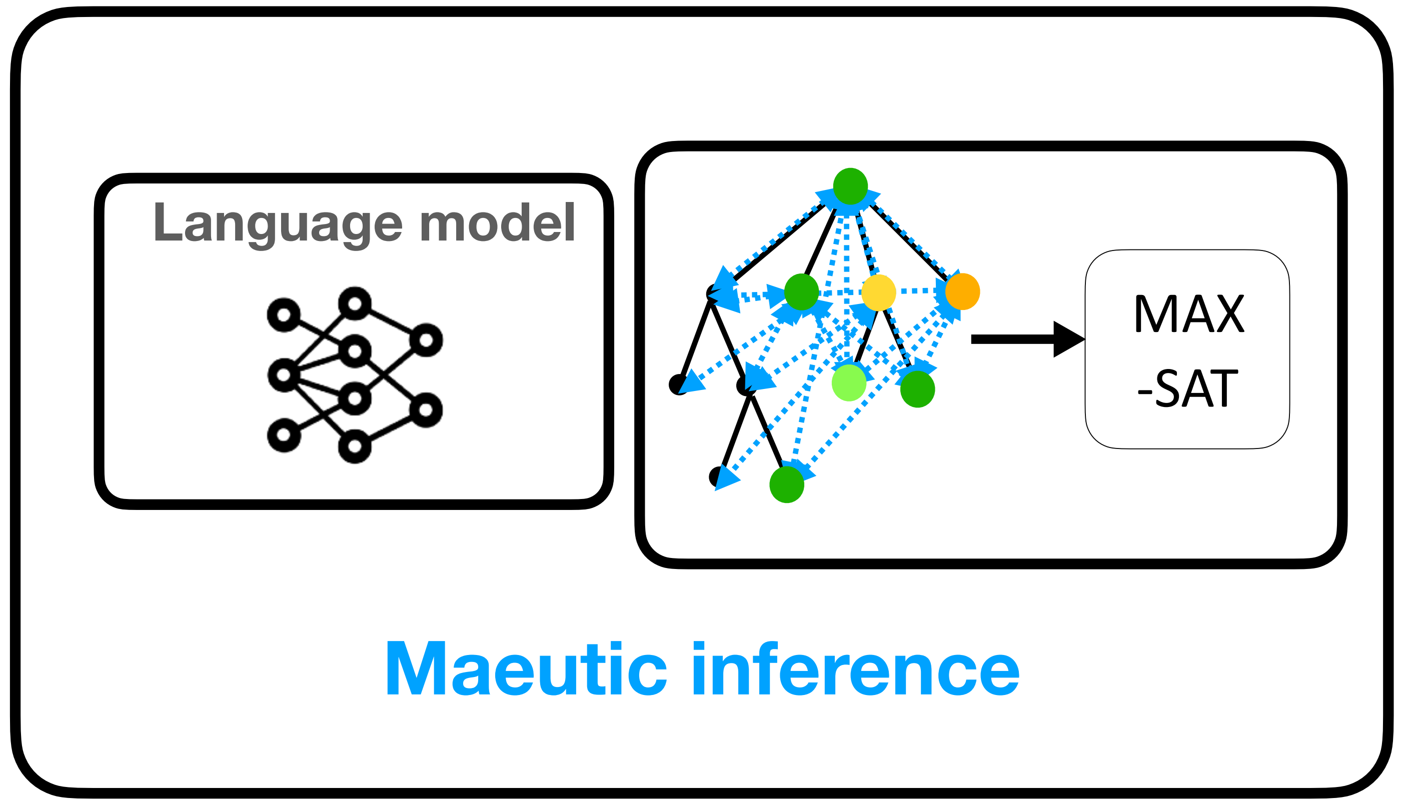
Summary

- Maeutic inference:
  - Recursively enumerate propositions
  - Assign confidence and identify contradictions
  - Globally resolve into a decision



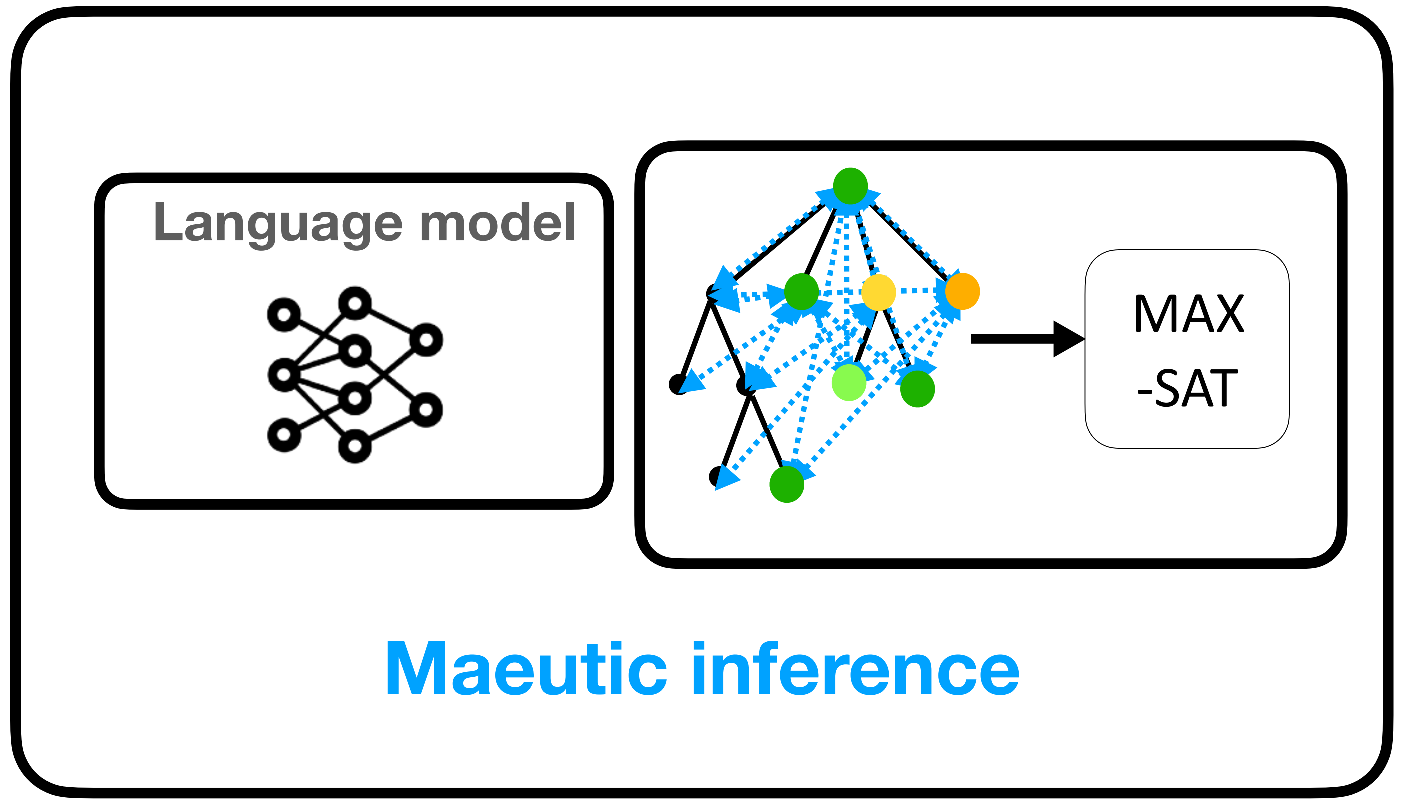
Summary

- Maeutic inference:
  - Recursively enumerate propositions
  - Assign confidence and identify contradictions
  - Globally resolve into a decision
- Strong off-the-shelf performance



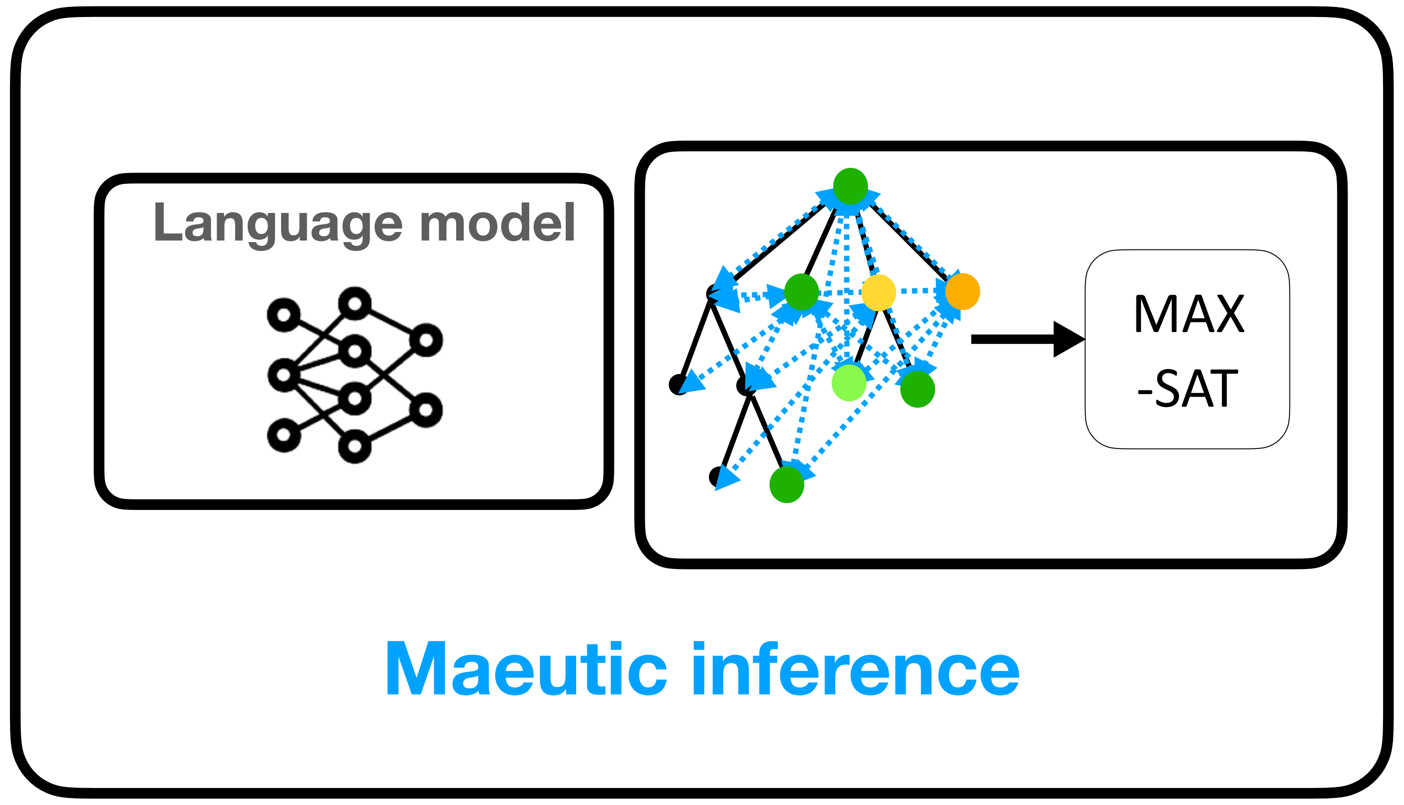
Summary

- Maeutic inference:
  - Recursively enumerate propositions
  - Assign confidence and identify contradictions
  - Globally resolve into a decision
- Strong off-the-shelf performance
- Interpretable interface



Summary

- Maeutic inference:
  - Recursively enumerate propositions
  - Assign confidence and identify contradictions
  - Globally resolve into a decision
- Strong off-the-shelf performance
- Interpretable interface
- Next steps: more complex label space, other creative algorithms



Thank you!

Led by:
Jaehun Jung

Lianhui Qin

Faeze
Brahman

Chandra
Bhagavatula

Ronan
Le Bras

Yejin Choi

Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations
https://arxiv.org/abs/2205.11822
Under Review