# Generating Natural Language Proofs with Verifier-Guided Search

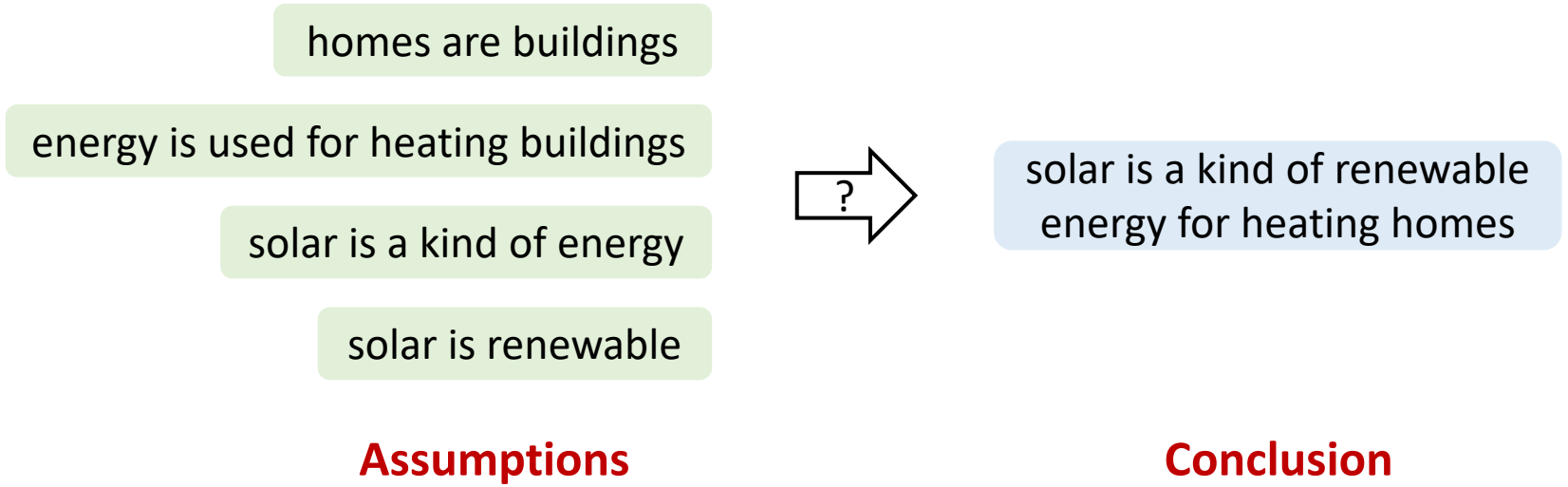
Kaiyu Yang, Jia Deng, Danqi Chen



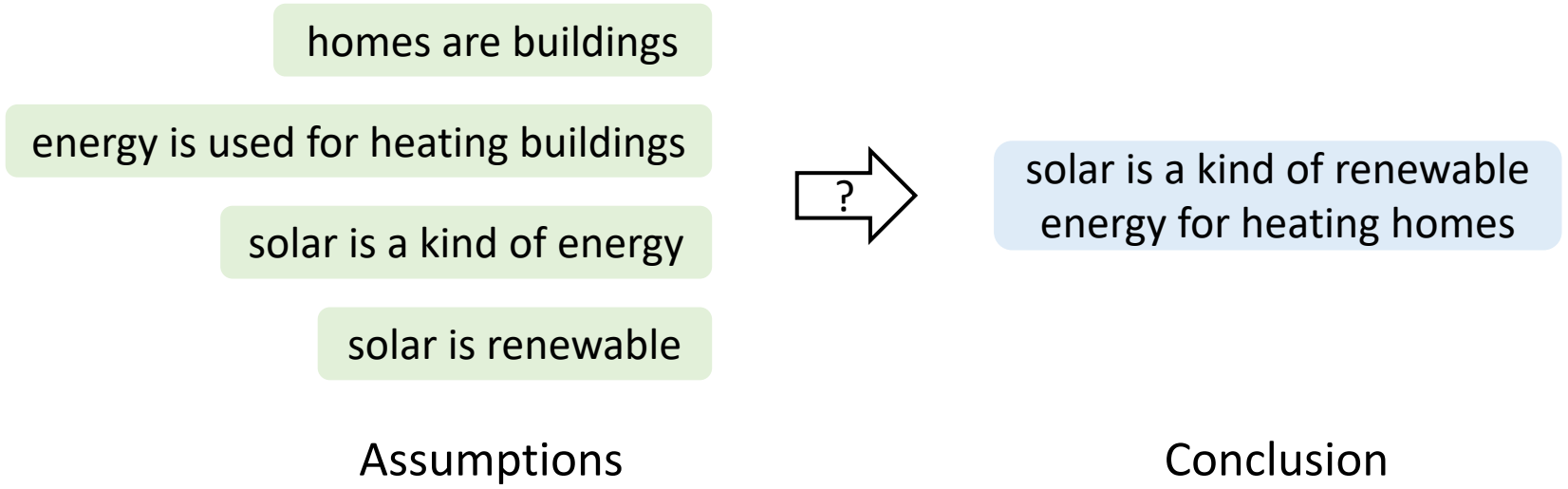
PRINCETON UNIVERSITY

PRINCETON NLP

# Reasoning in Natural Language



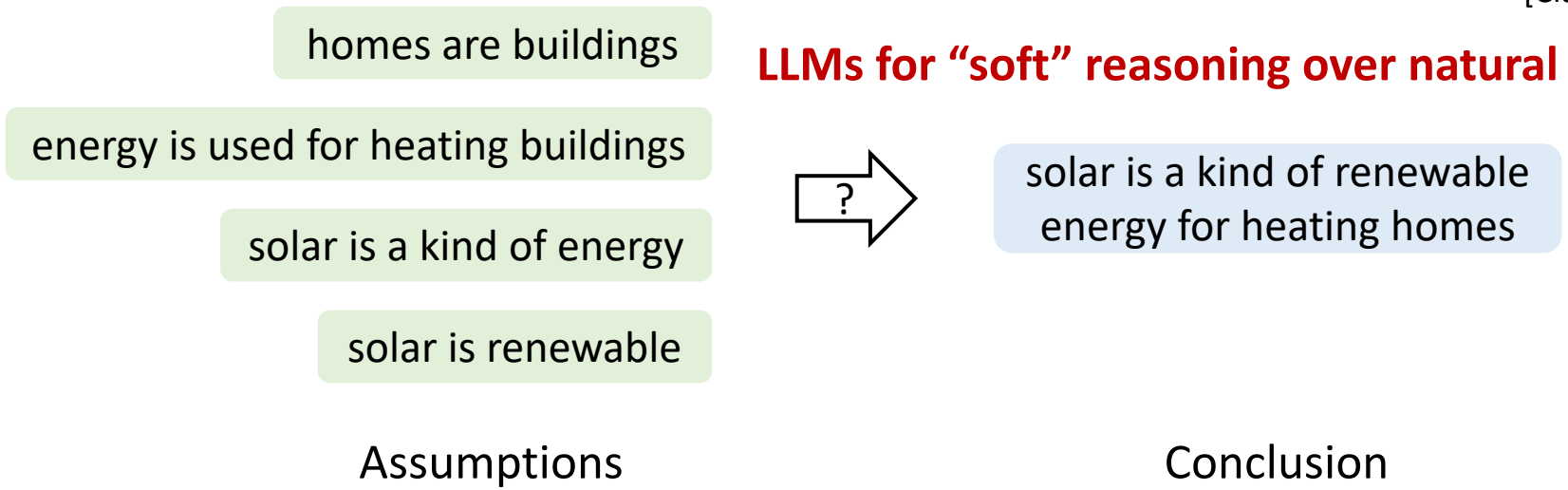
# Reasoning in Natural Language




- Studied extensively in automated theorem proving
- **Remains challenging in natural language**
  - Fuzzy, imprecise, requiring implicit knowledge
  - No well-defined inference rules

# Reasoning in Natural Language

[Clark et al. IJCAI 2020]




**LLMs for "soft" reasoning over natural language**

- Studied extensively in automated theorem proving
- Remains challenging in natural language
  - Fuzzy, imprecise, requiring implicit knowledge
  - No well-defined inference rules

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

**Supporting facts ($C$):**

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

**Hypothesis ($h$):**

*h*: solar is a kind of renewable energy for heating homes

**Supporting facts ($C$):**

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

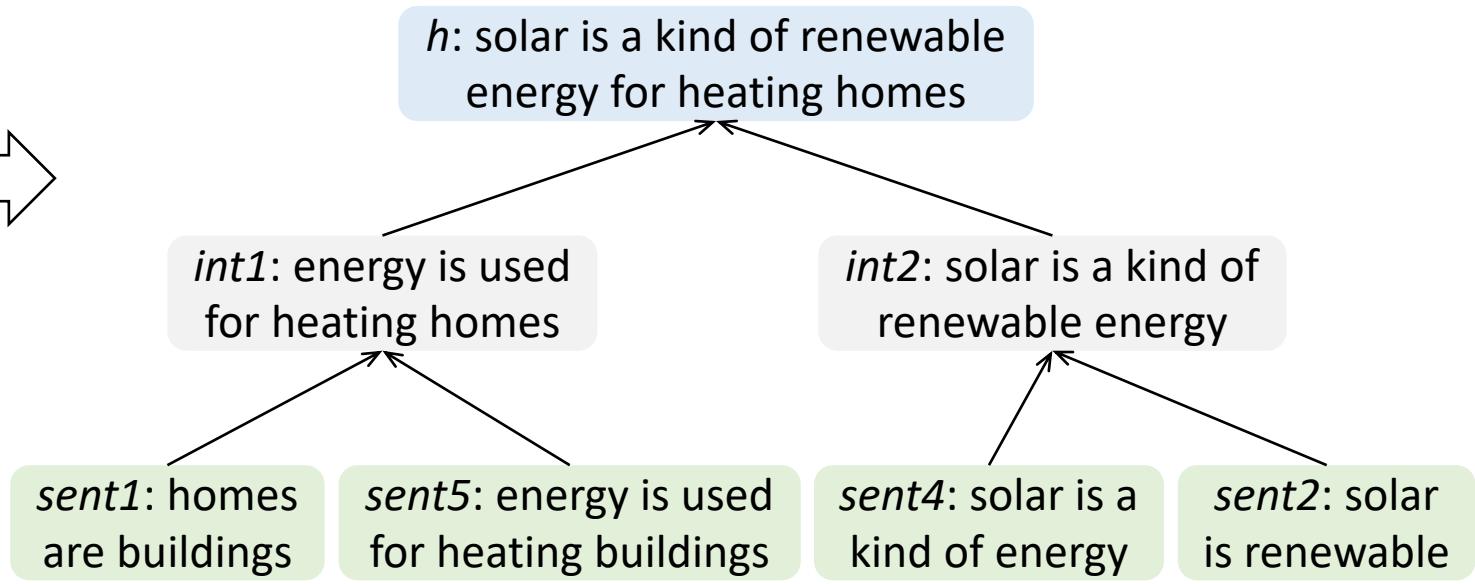
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

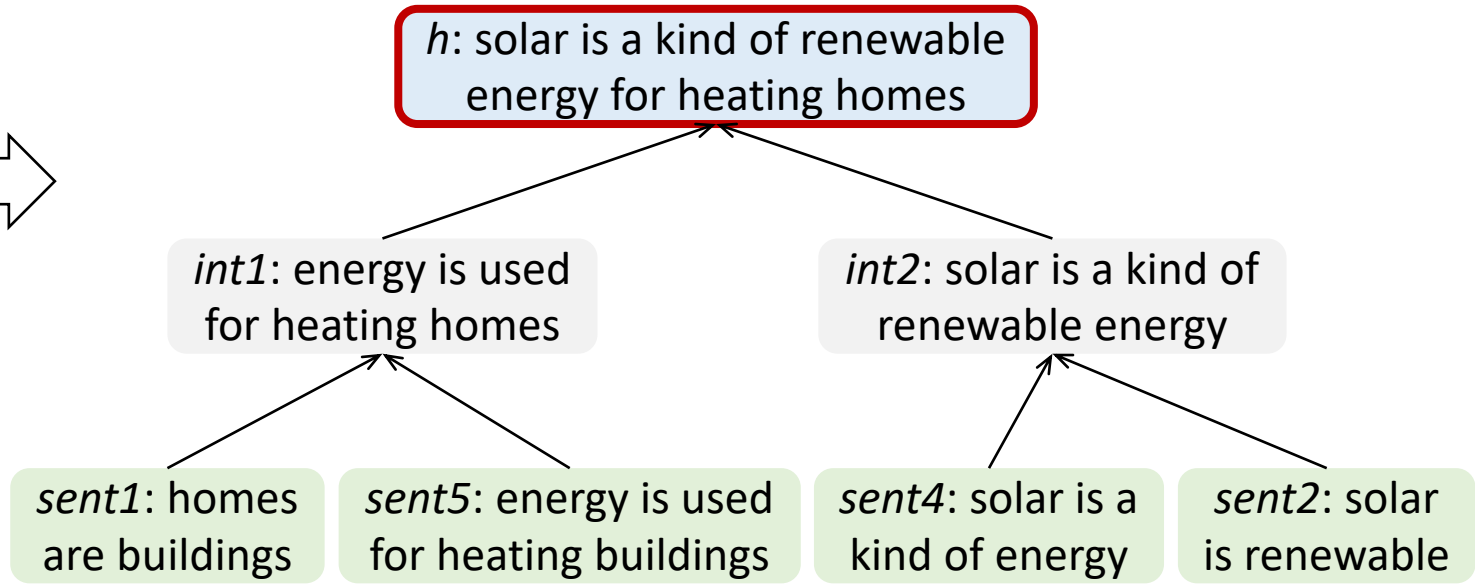
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

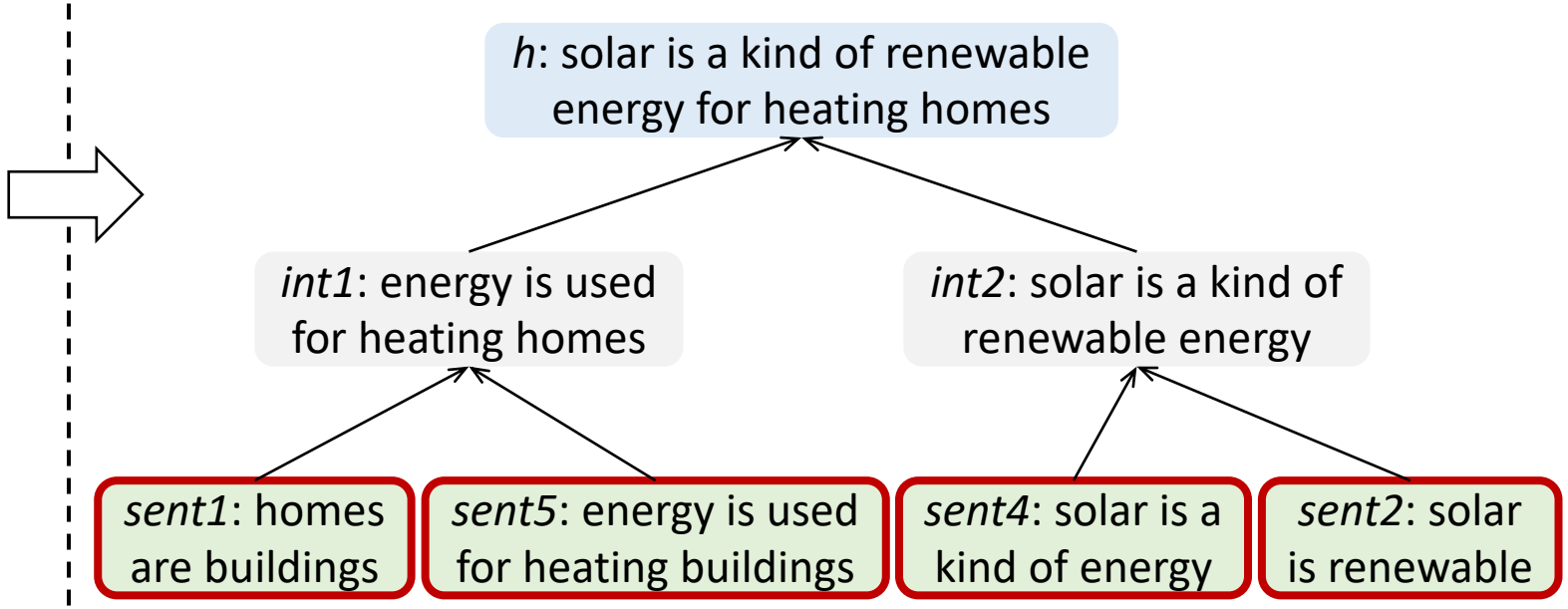
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Task: Proof Generation

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

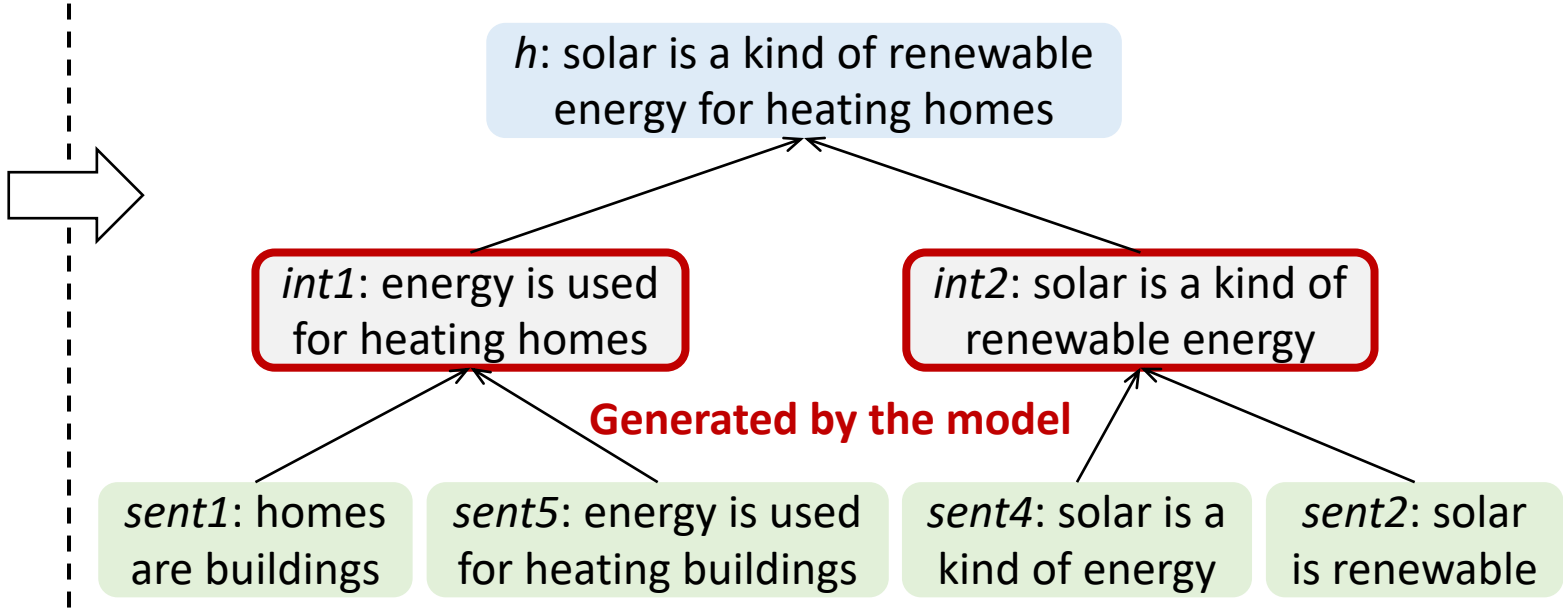
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Single-Shot Methods

**Generate the entire proof altogether**

Hypothesis ($h$):

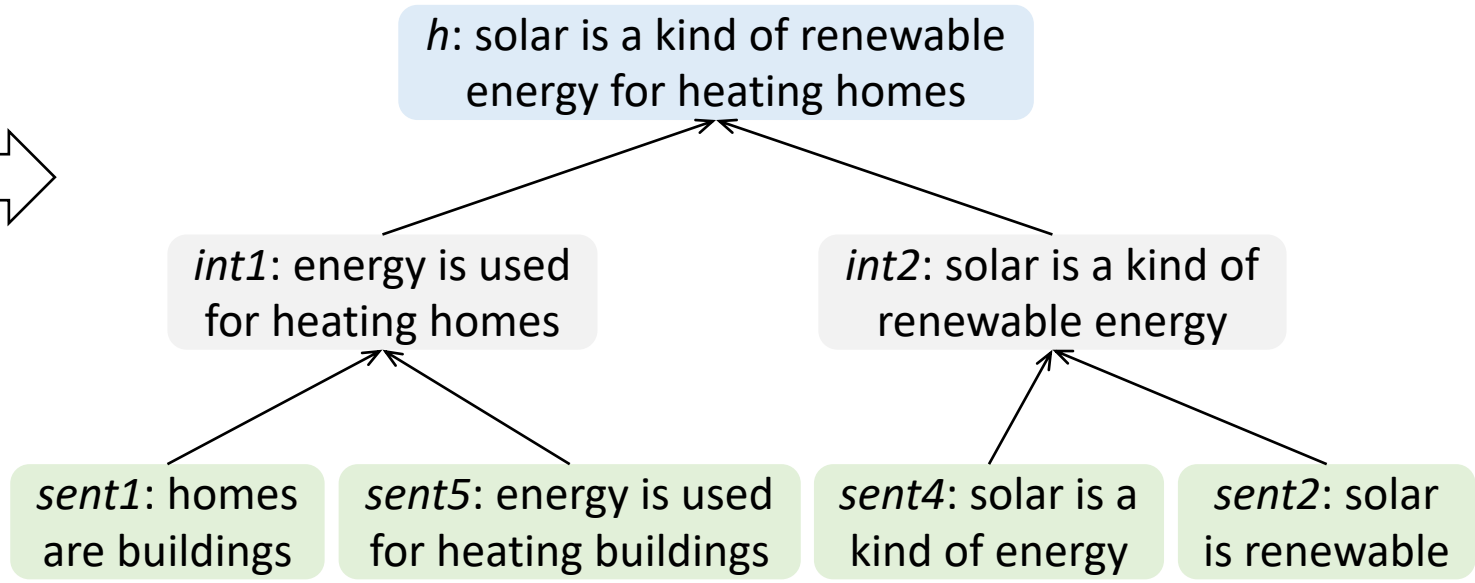
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Stepwise Methods

**Generate the proof step by step**

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

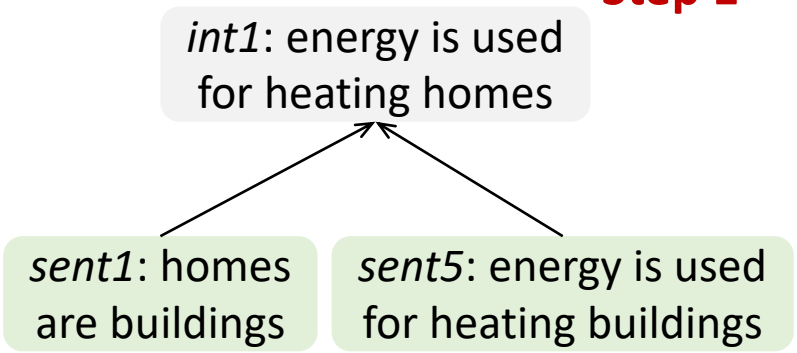
Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):

**Step 1**




**Output**

# Stepwise Methods

**Generate the proof step by step**

Hypothesis ($h$):

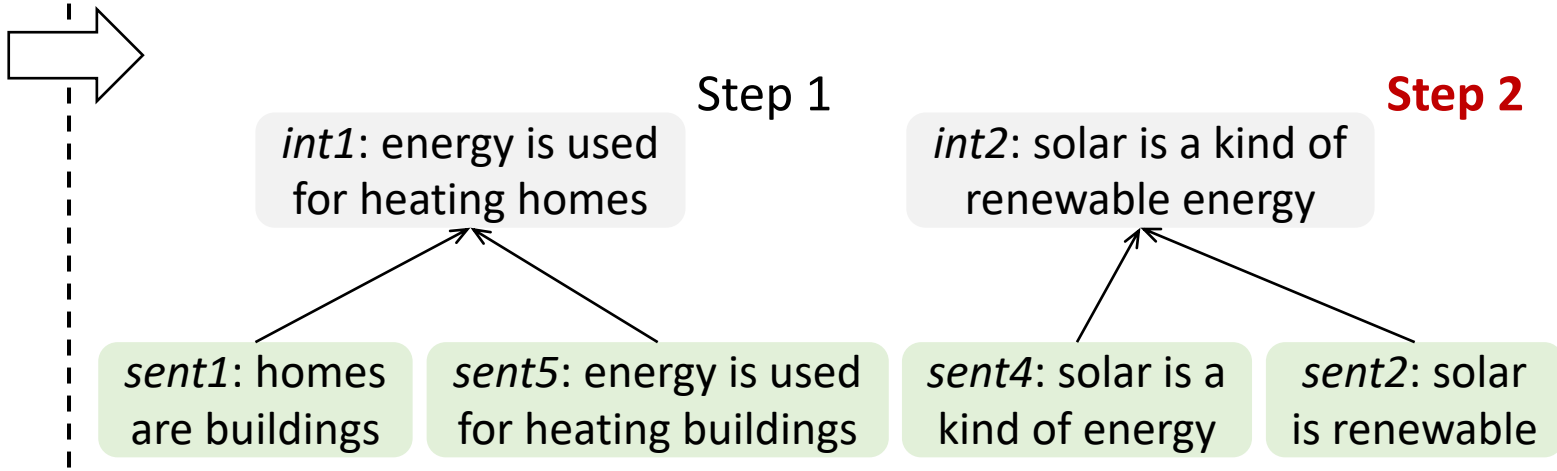
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Stepwise Methods

**Generate the proof step by step**

Hypothesis ($h$):

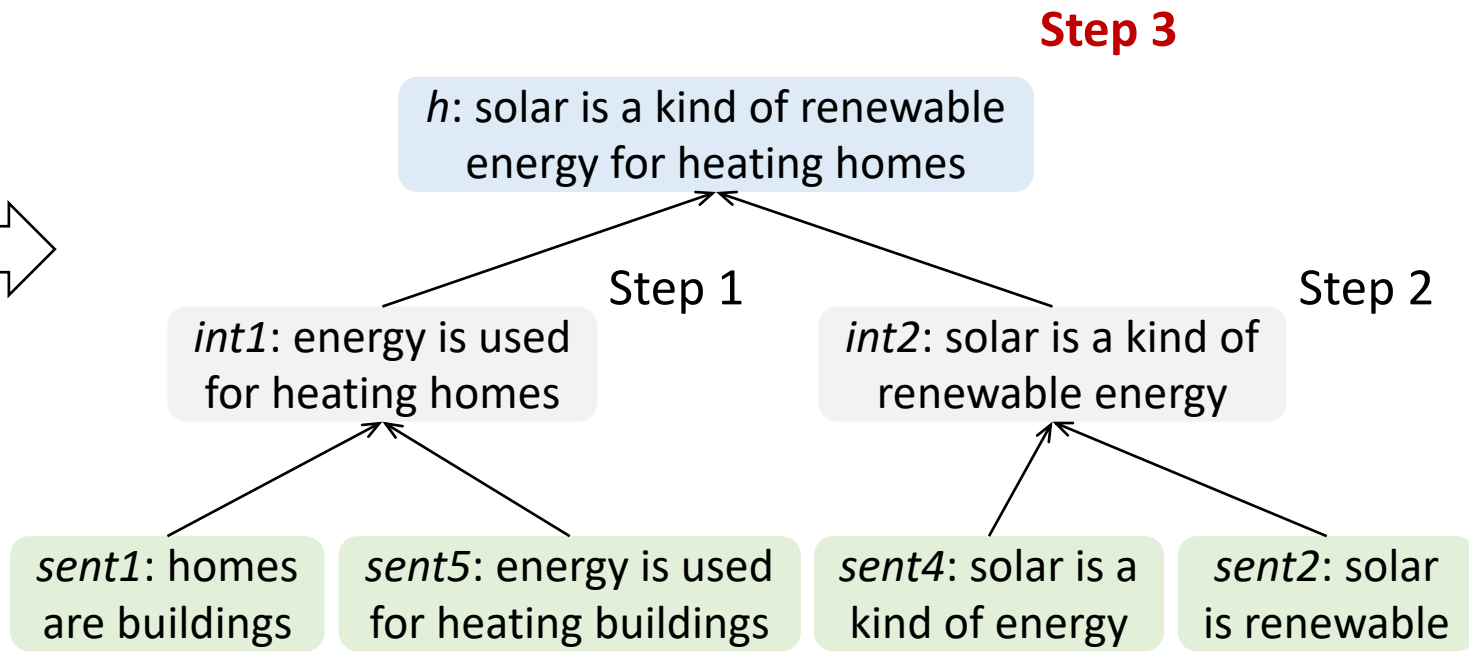
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

Proof tree ($T$):




**Output**

# Single-Shot vs. Stepwise Methods

**Generate the entire proof altogether**

- PRover [Saha et al. EMNLP 2020]
- EntailmentWriter [Dalvi et al. EMNLP 2021]
- PRobr [Sun et al. Findings of ACL 2021]

**Generate the proof step by step**

- ProofWriter [Tafjord et al. Findings of ACL 2021]
- FaiRR [Sanyal et al. ACL 2022]
- SCSearch [Bostrom et al. Findings of EMNLP 2022]
- MetGen [Hong et al. Findings of NAACL 2022]

Can better leverage compositionality and generalize to longer proofs

Achieved limited success on challenging proofs authored by humans (e.g., EntailmentBank)

# Stepwise Methods

[Dalvi et al. EMNLP 2021]

Hypothesis ($h$):

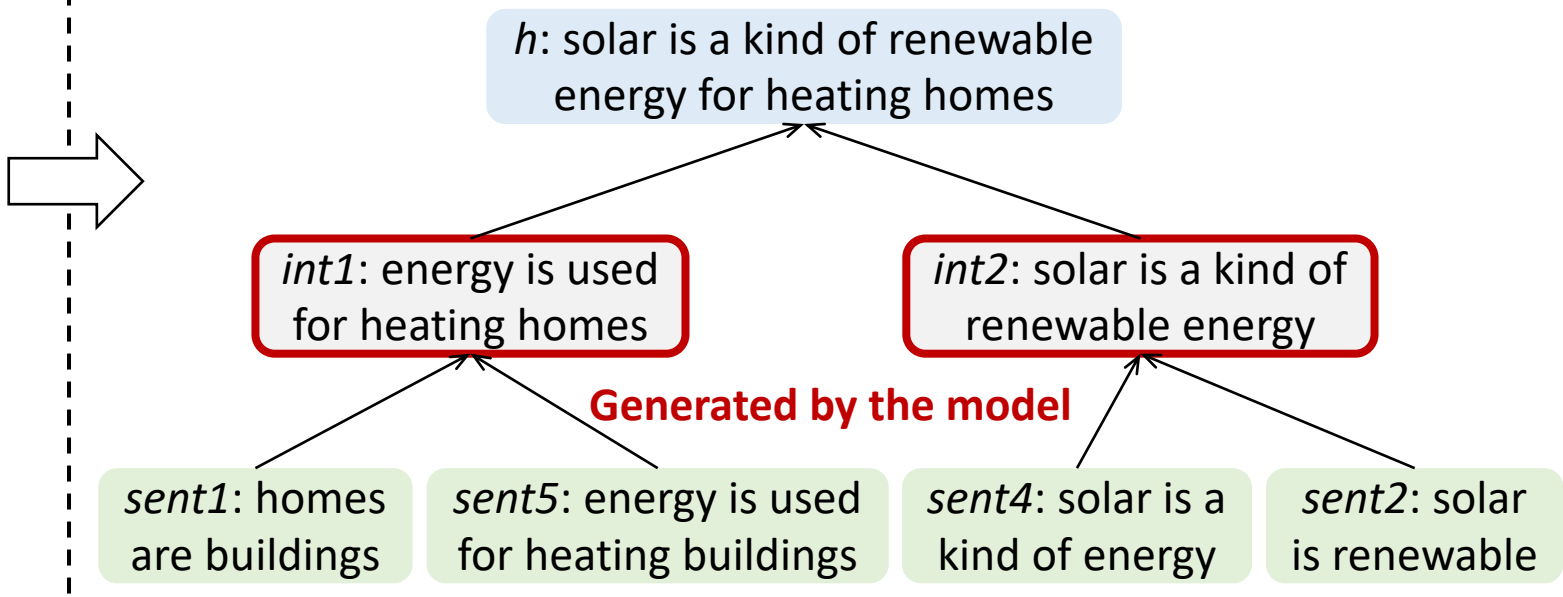
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

**Input**

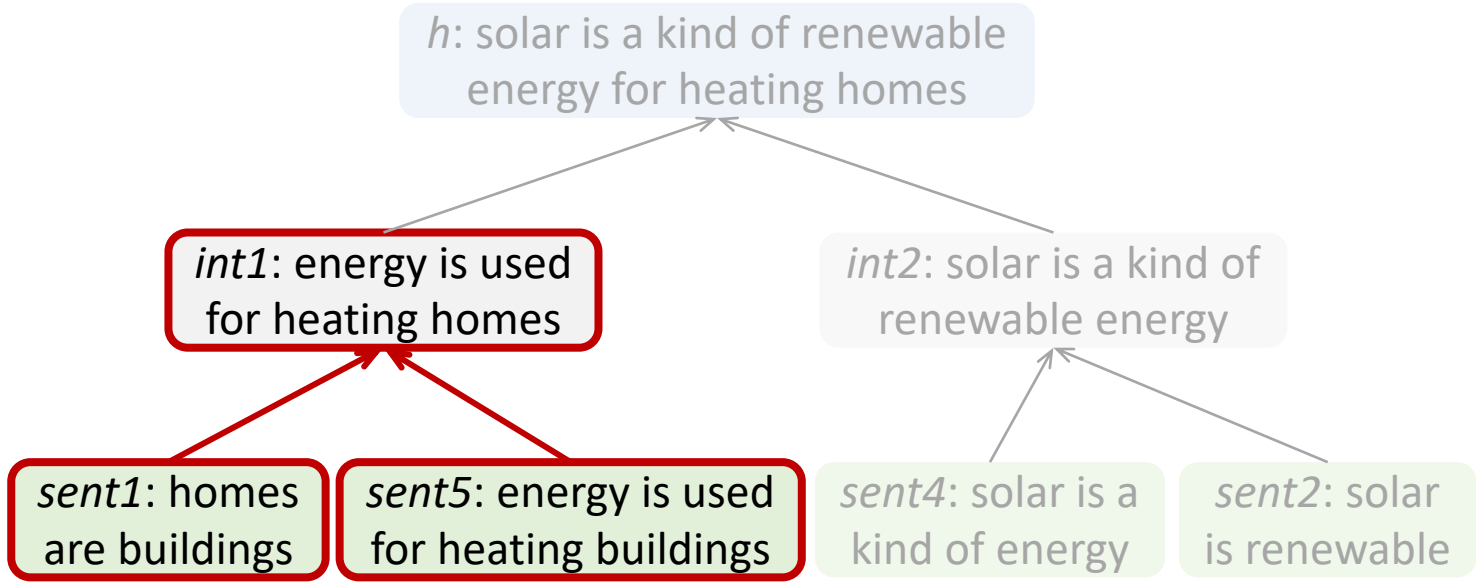
Proof tree ($T$):




**Output**

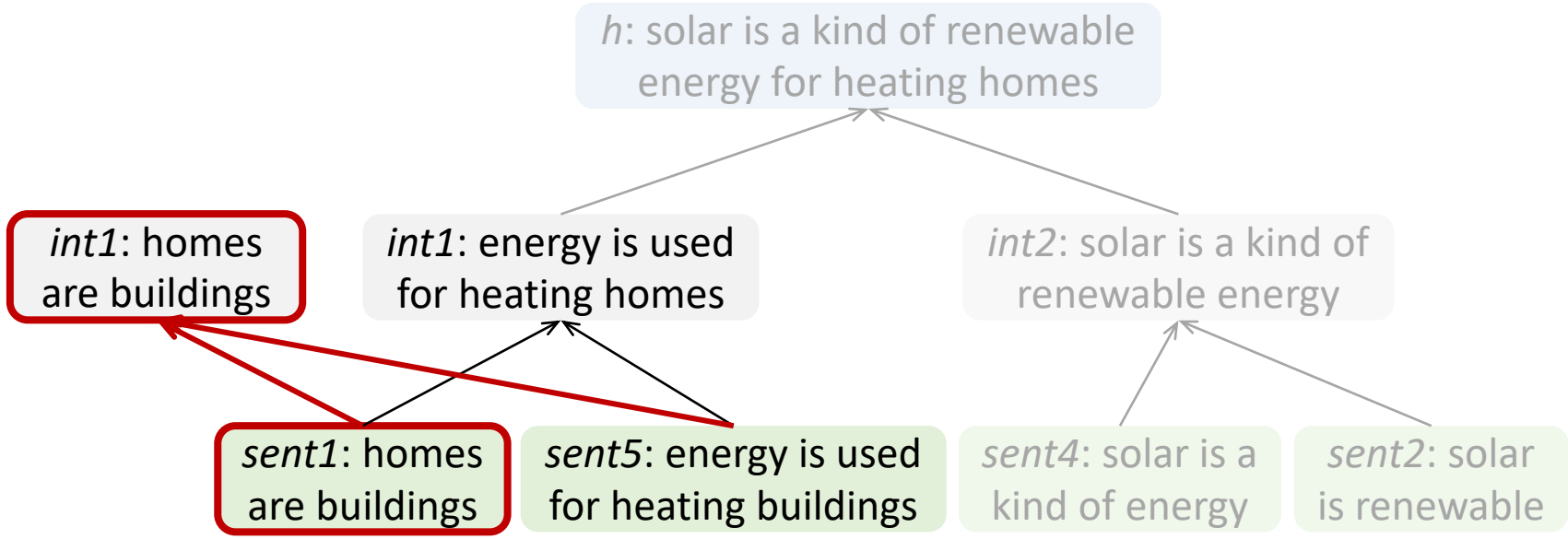
# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)



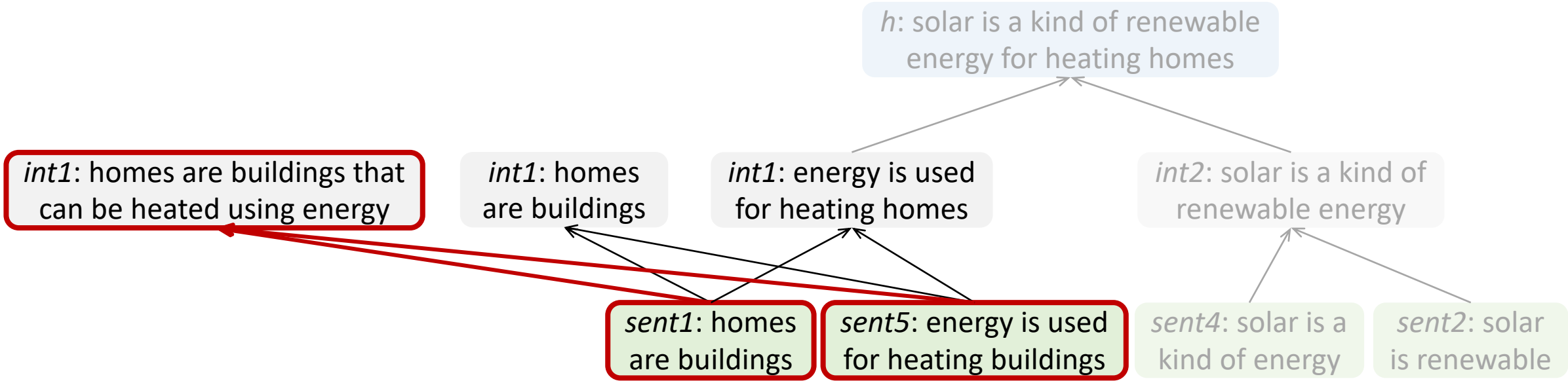
# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)



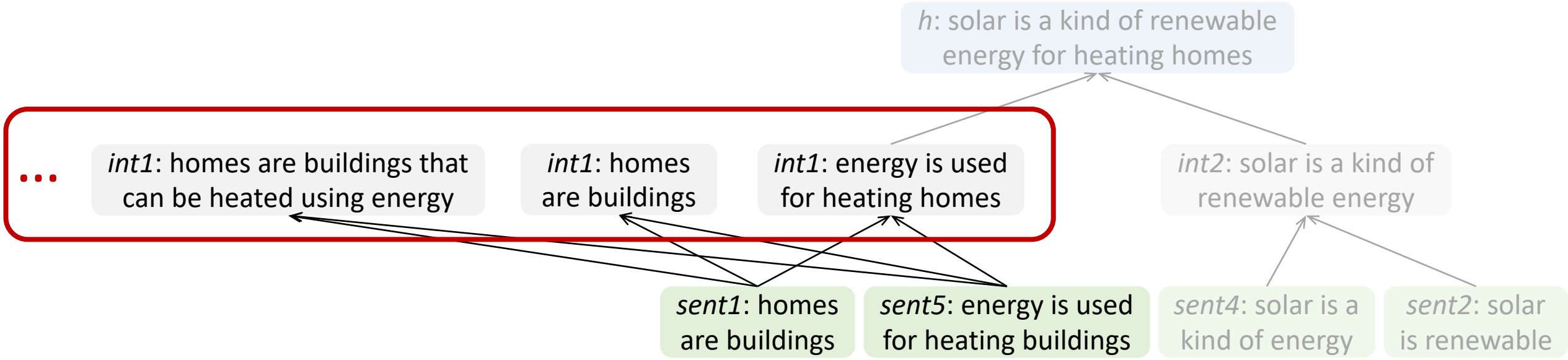
# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)



# Challenges in Generating Valid and Relevant Steps

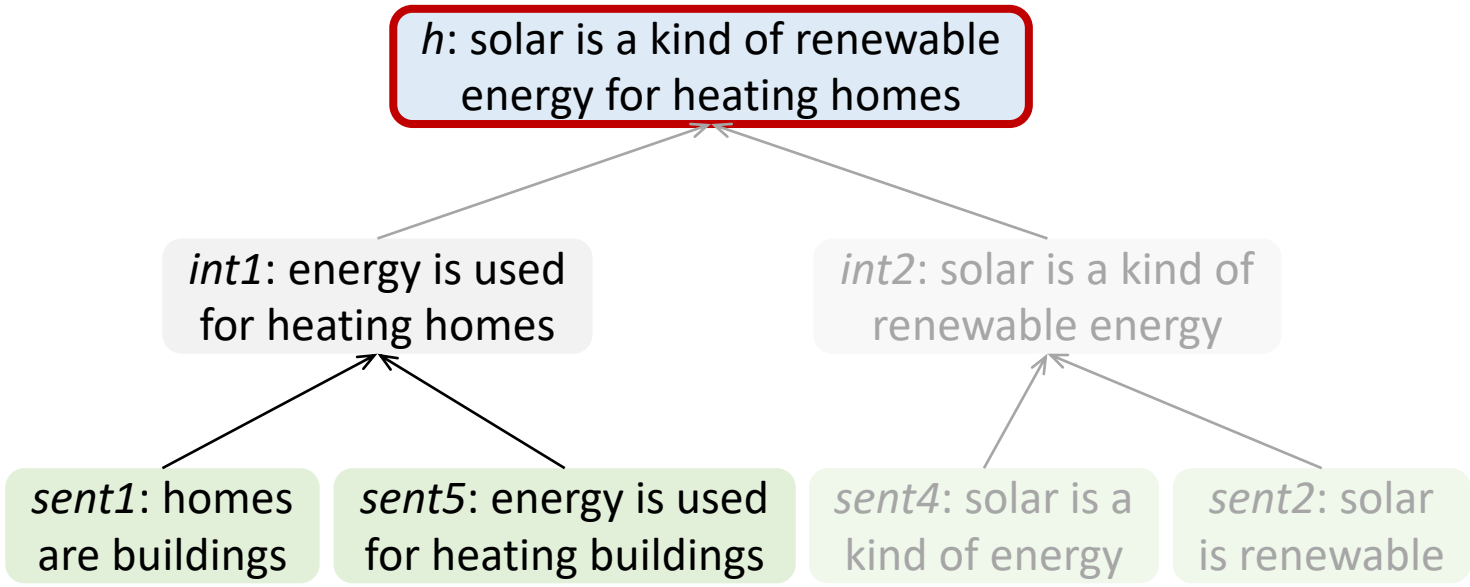
- Many valid steps are irrelevant (not useful for proving the hypothesis)



# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

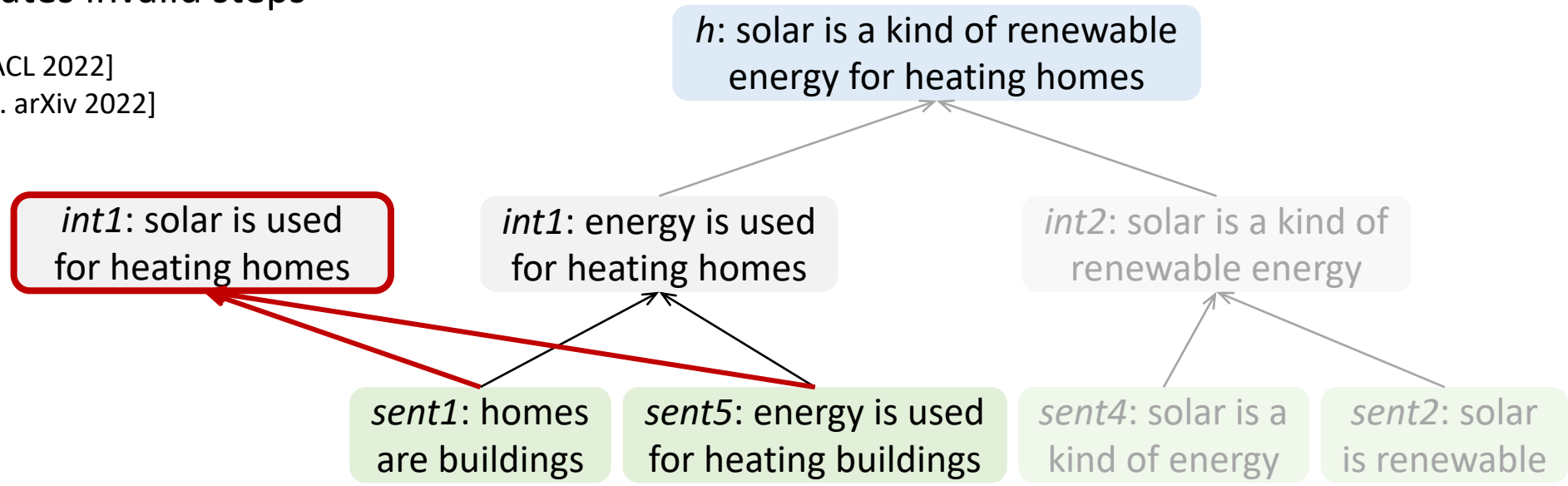
[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]



# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

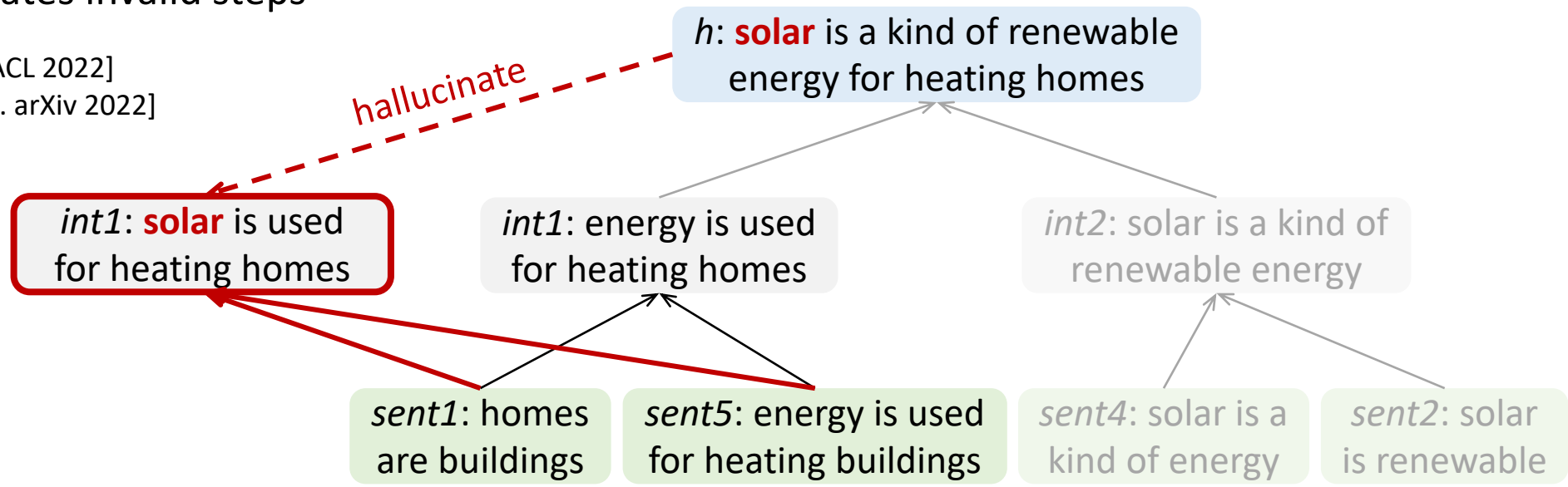
[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]



# Challenges in Generating Valid and Relevant Steps

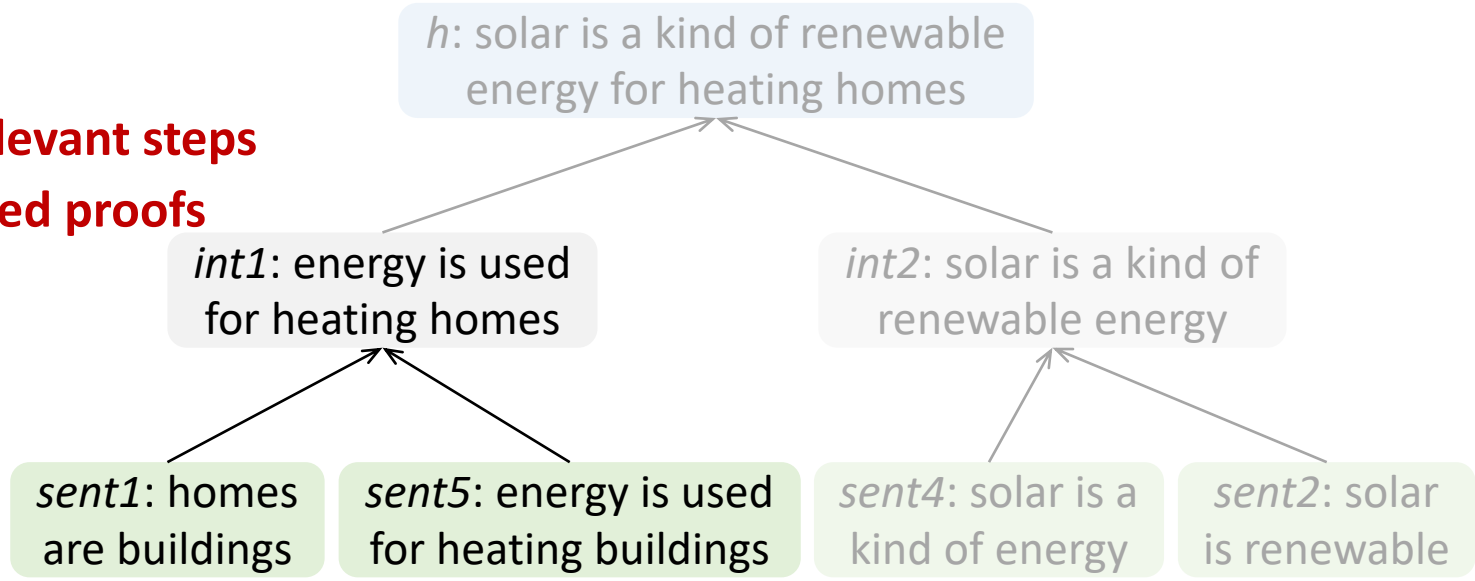
- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps

[Sanyal et al. ACL 2022]
[Bostrom et al. arXiv 2022]



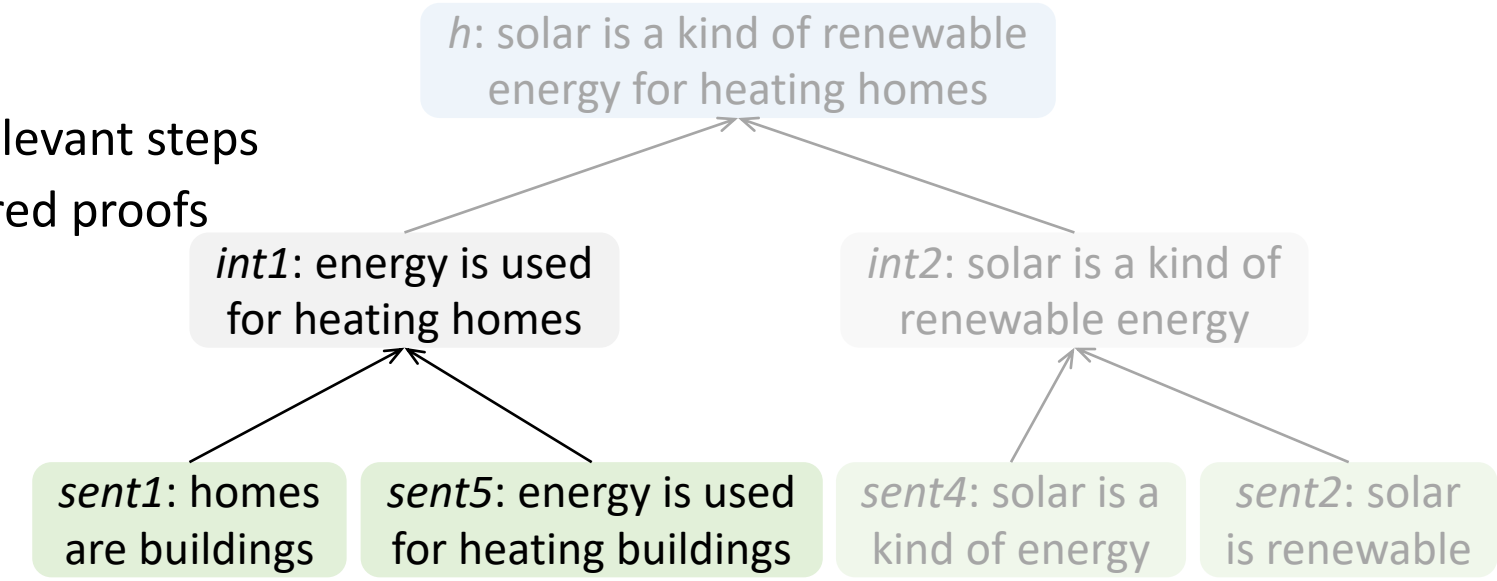
# Challenges in Generating Valid and Relevant Steps

- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps
- Existing stepwise methods
  - **Struggle to generate valid and relevant steps**
  - **Underperform on human-authored proofs**



# Challenges in Generating Valid and Relevant Steps

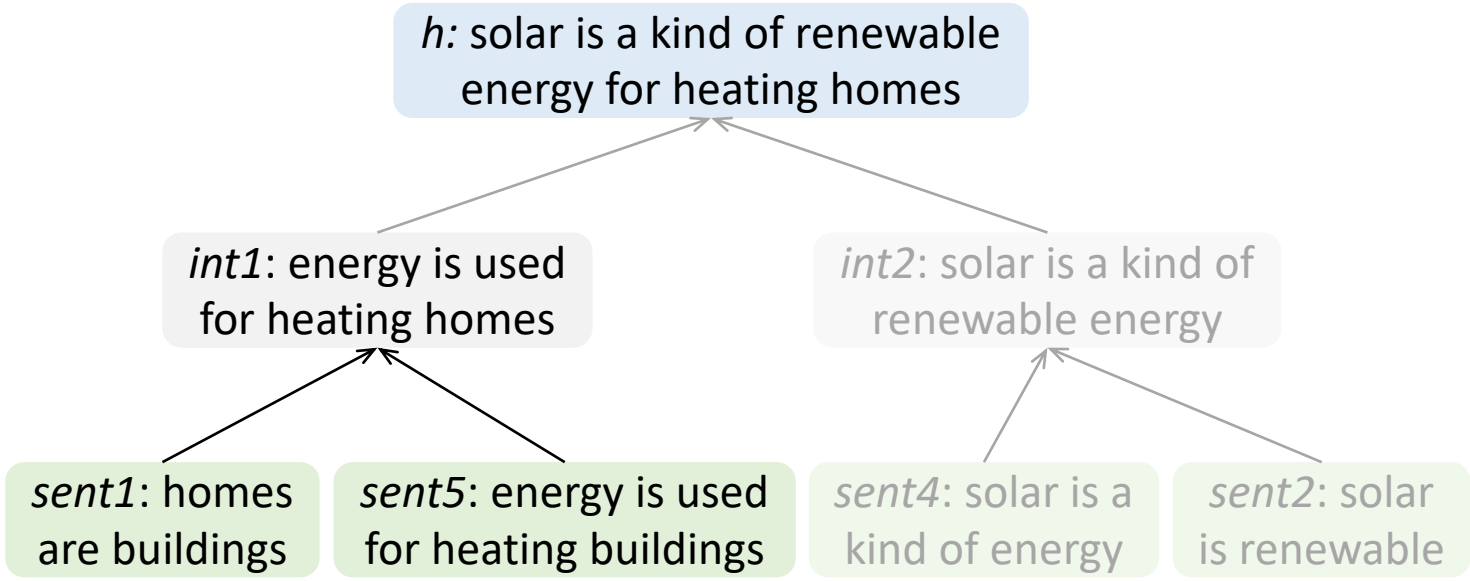
- Many valid steps are irrelevant (not useful for proving the hypothesis)
- The model hallucinates invalid steps
- Existing stepwise methods
  - Struggle to generate valid and relevant steps
  - Underperform on human-authored proofs




- **Our solution: a new method for stepwise proof generation**

# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**

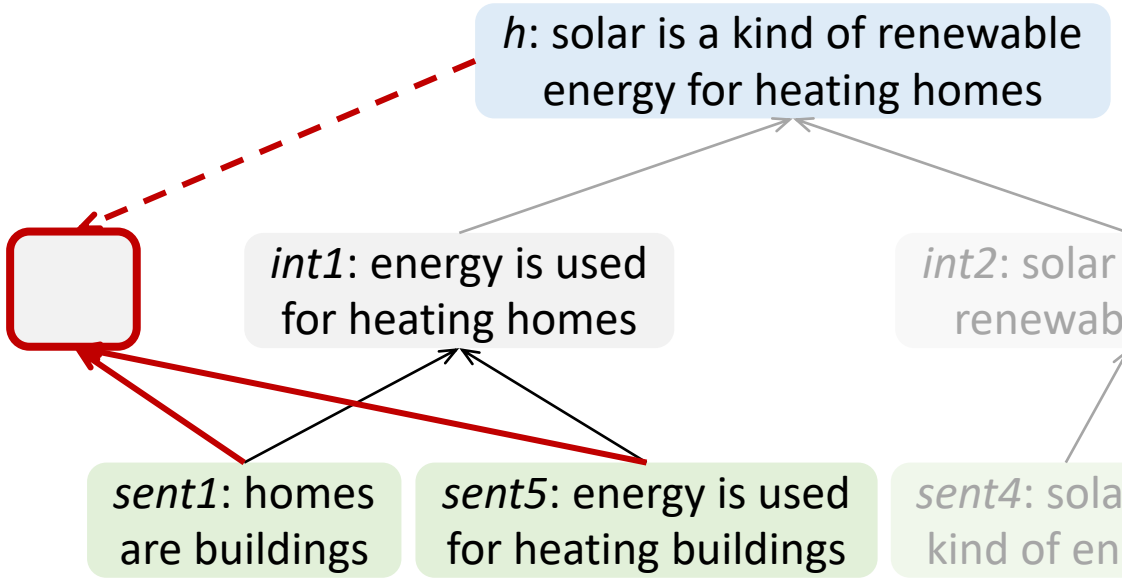


# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**

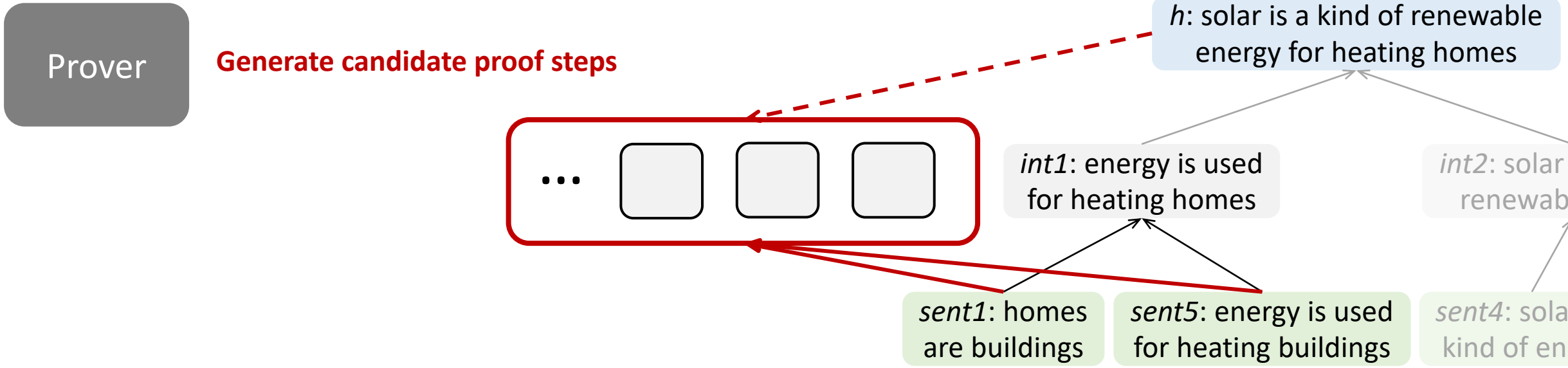
Prover

**Generate candidate proof steps**



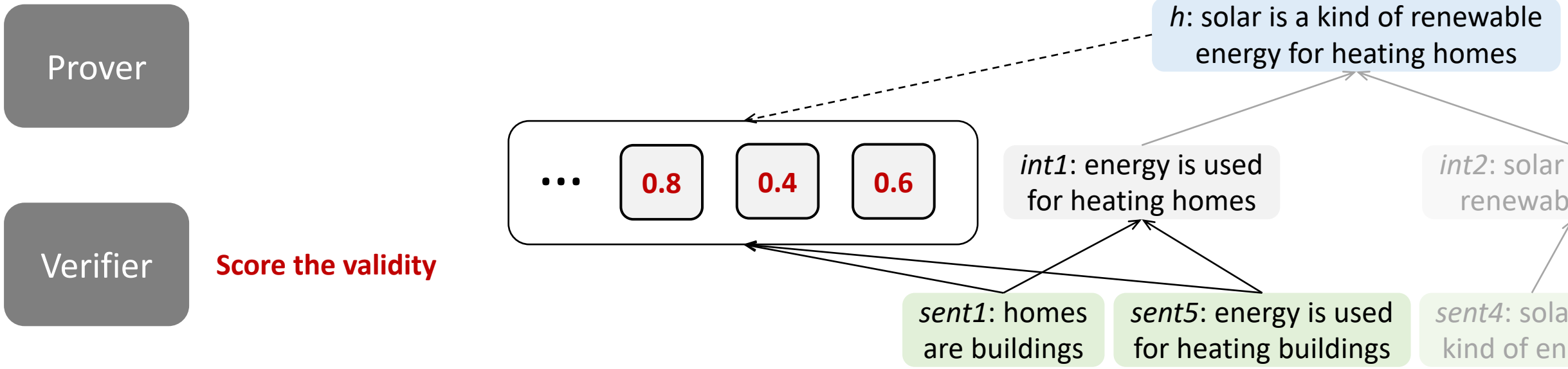
# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**



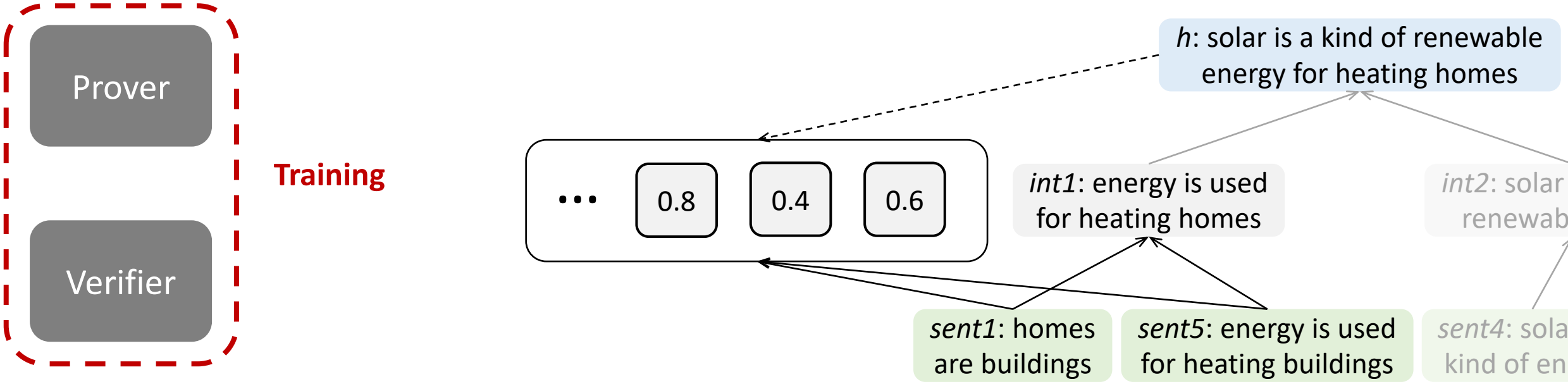
# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**



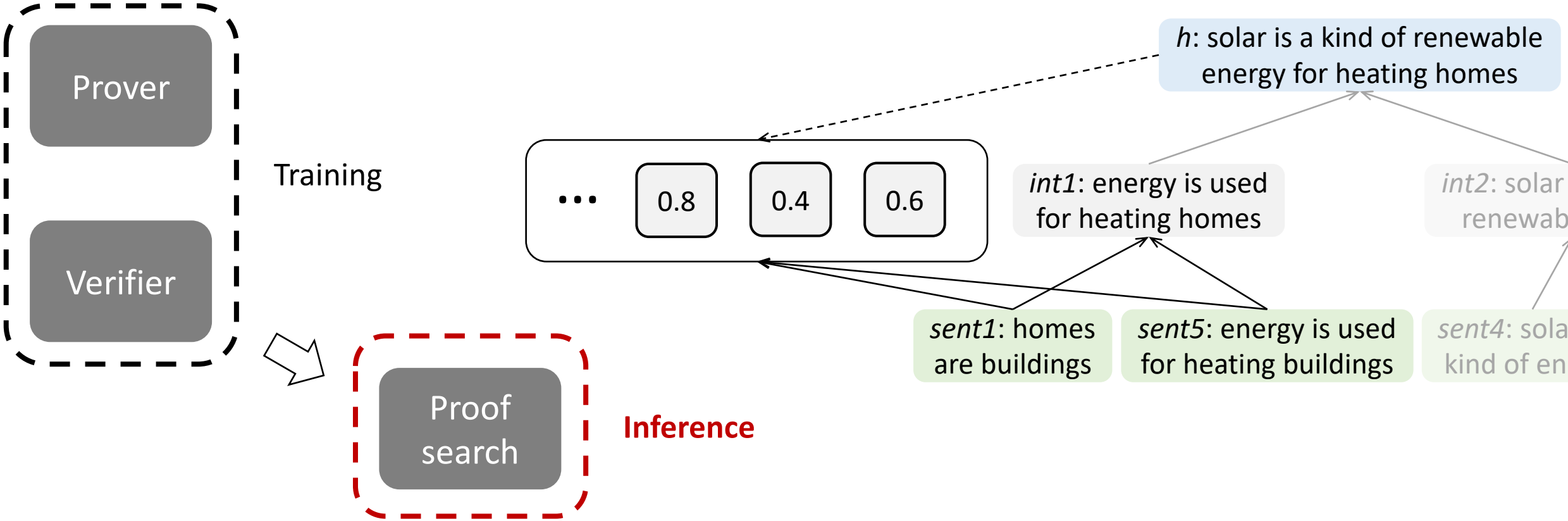
# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**



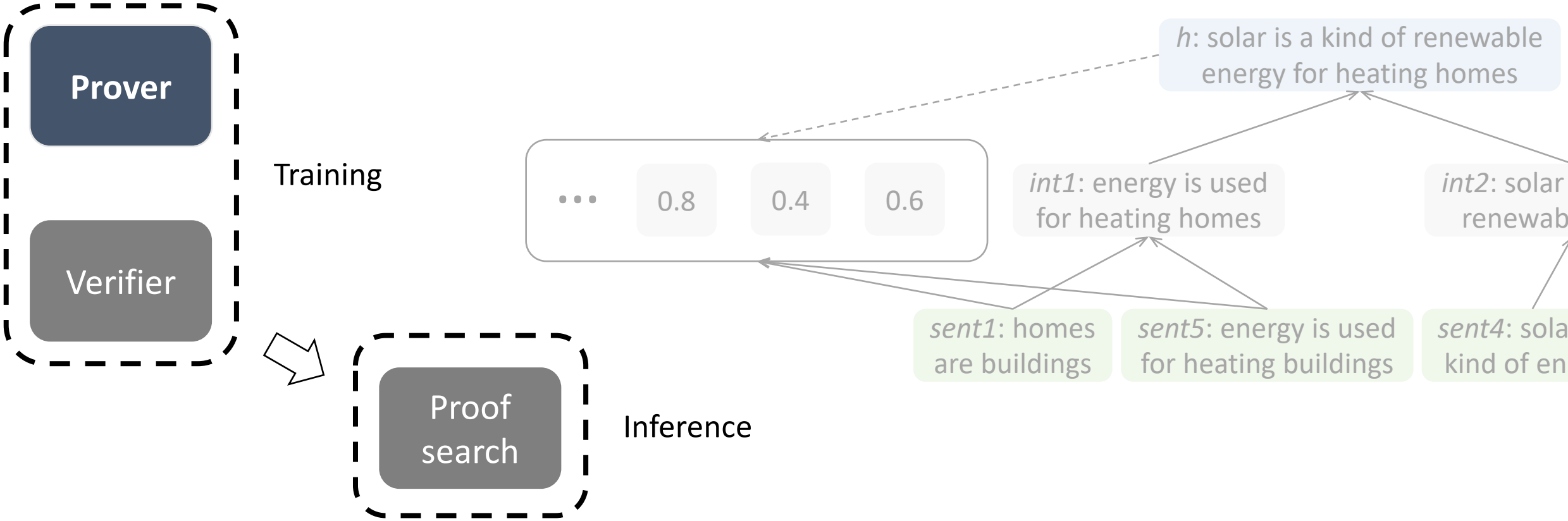
# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**



# NLProofS: Natural Language Proof Search

- A new method for **stepwise proof generation**



# Stepwise Prover

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…

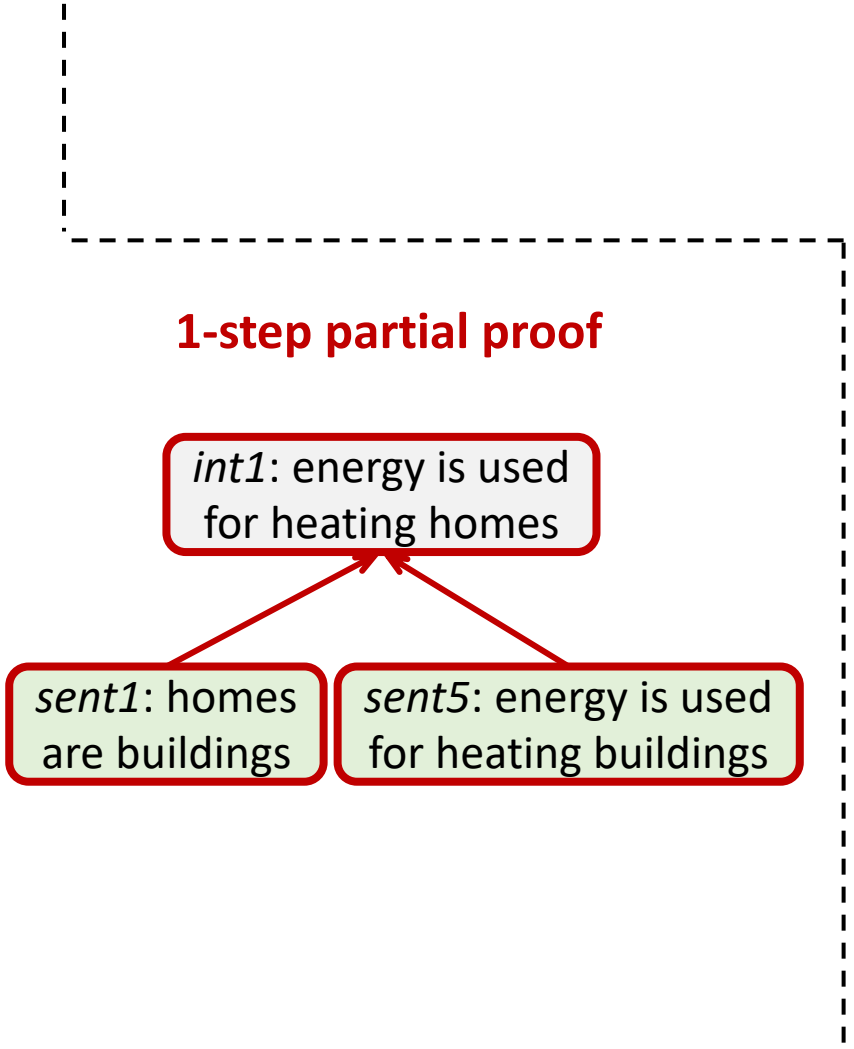
# Stepwise Prover

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…



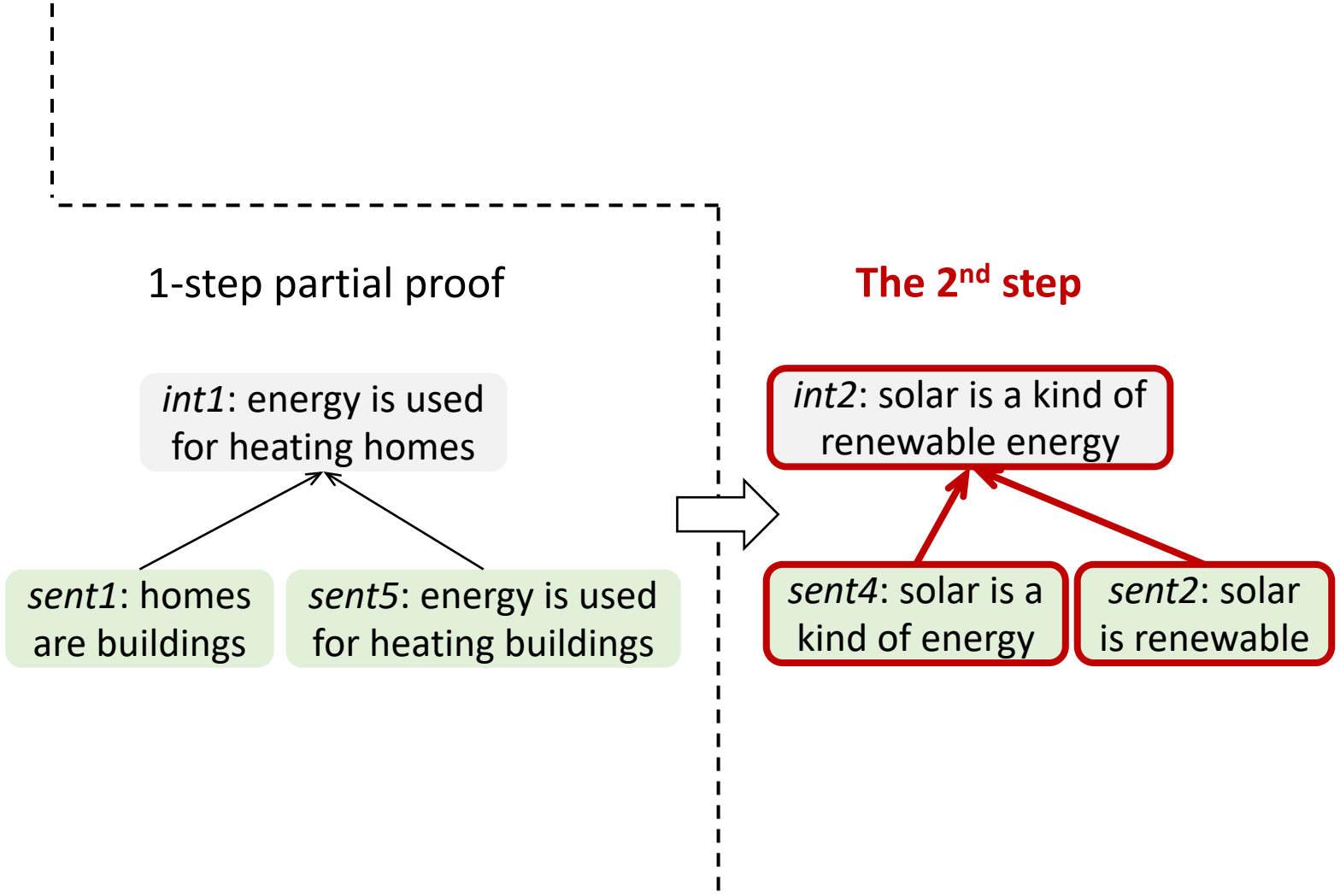
# Stepwise Prover

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…



# Stepwise Prover

[Raffle et al. JMLR 2020]

[Tafjord et al. Findings of ACL 2021]

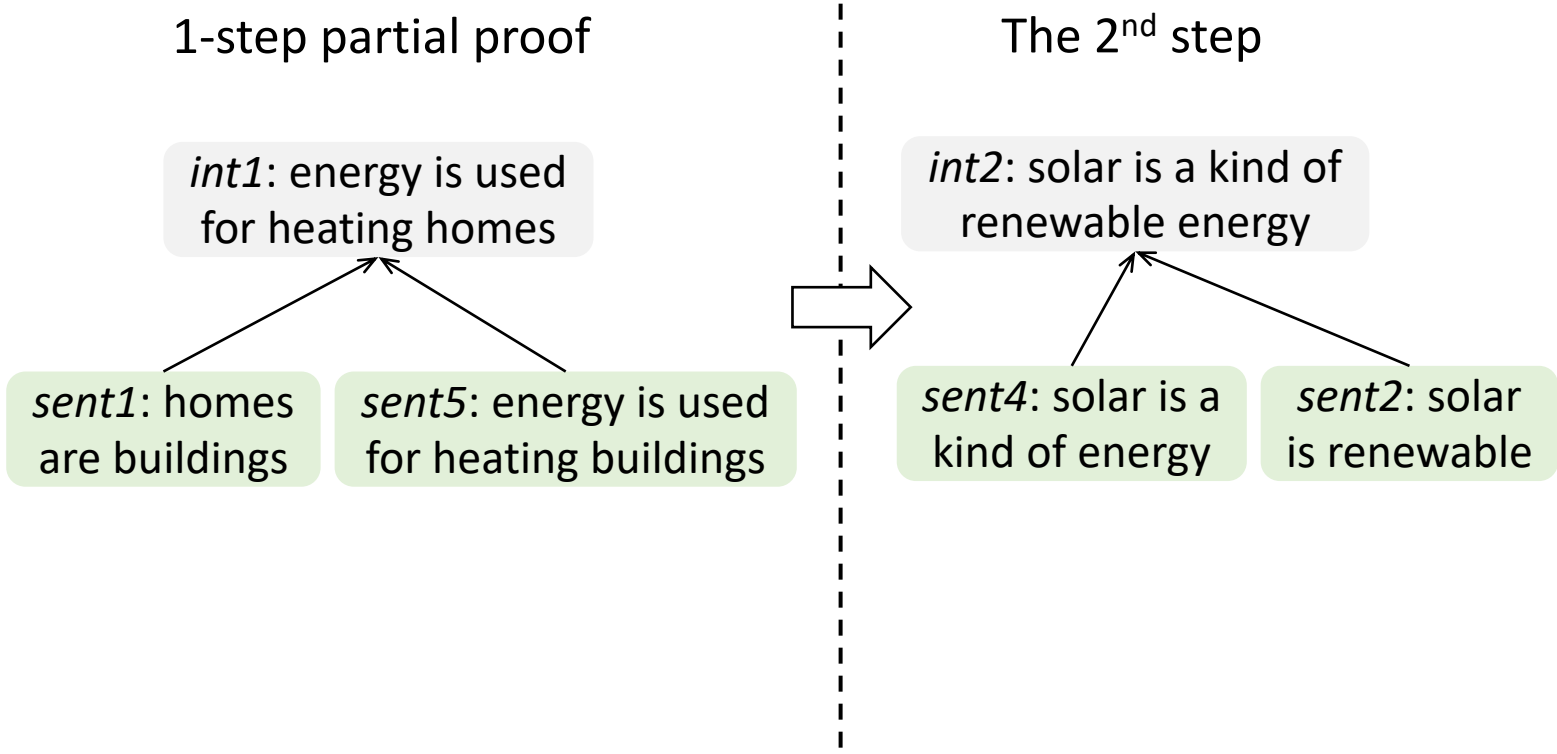
- **Finetune a T5 model to predict the next proof step**

Hypothesis ($h$):

*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…



# Stepwise Prover

[Raffle et al. JMLR 2020] [Tafjord et al. Findings of ACL 2021]

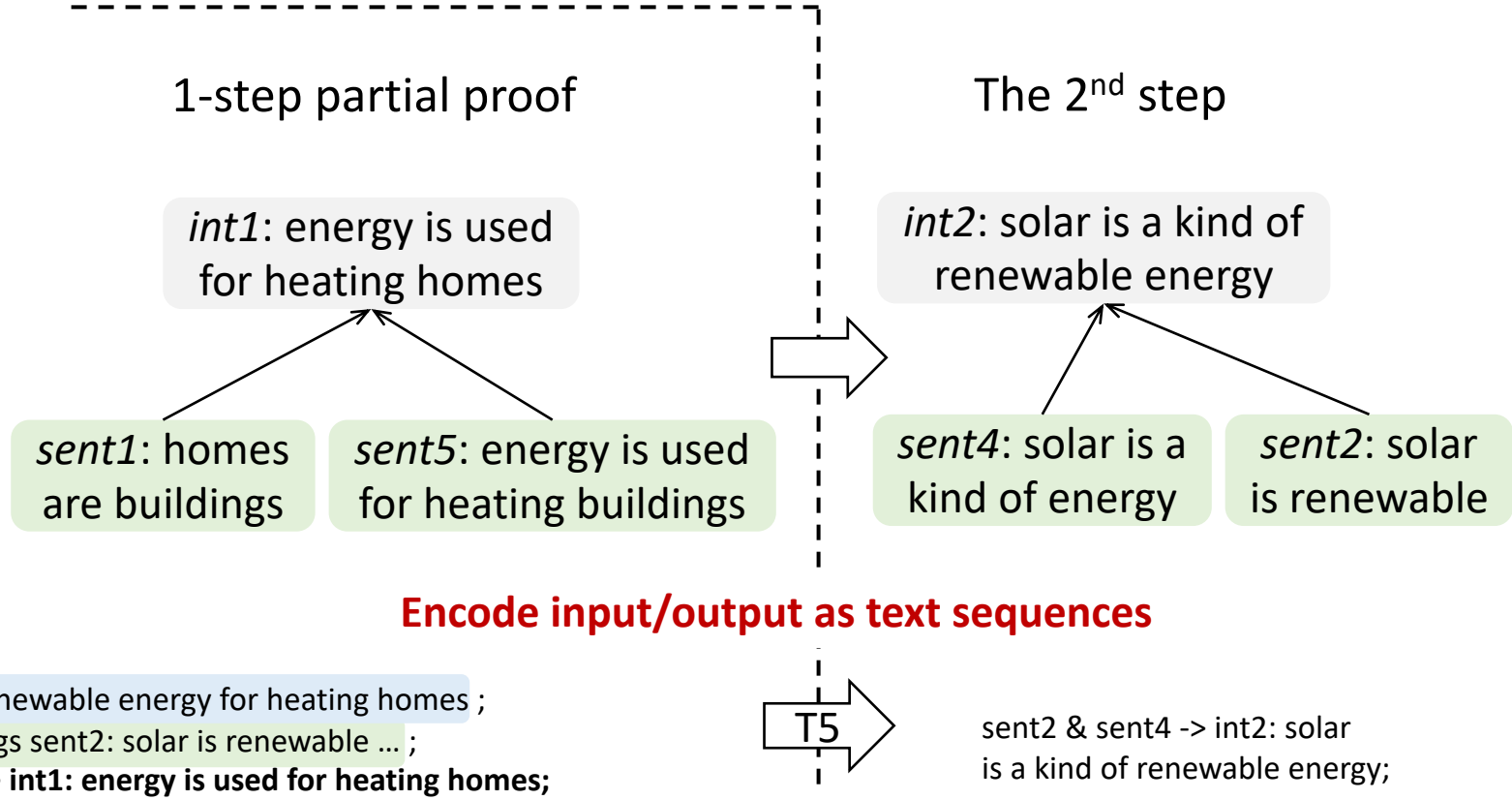
- Finetune a T5 model to predict the next proof step

Hypothesis ($h$):

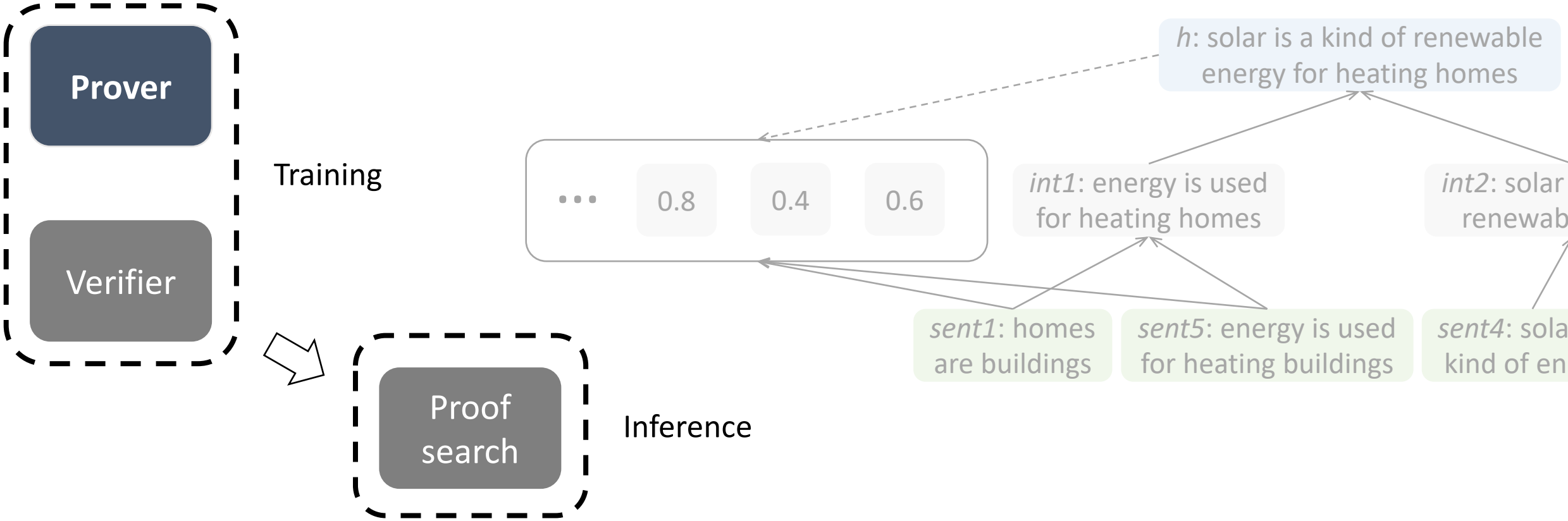
*h*: solar is a kind of renewable energy for heating homes

Supporting facts ($C$):

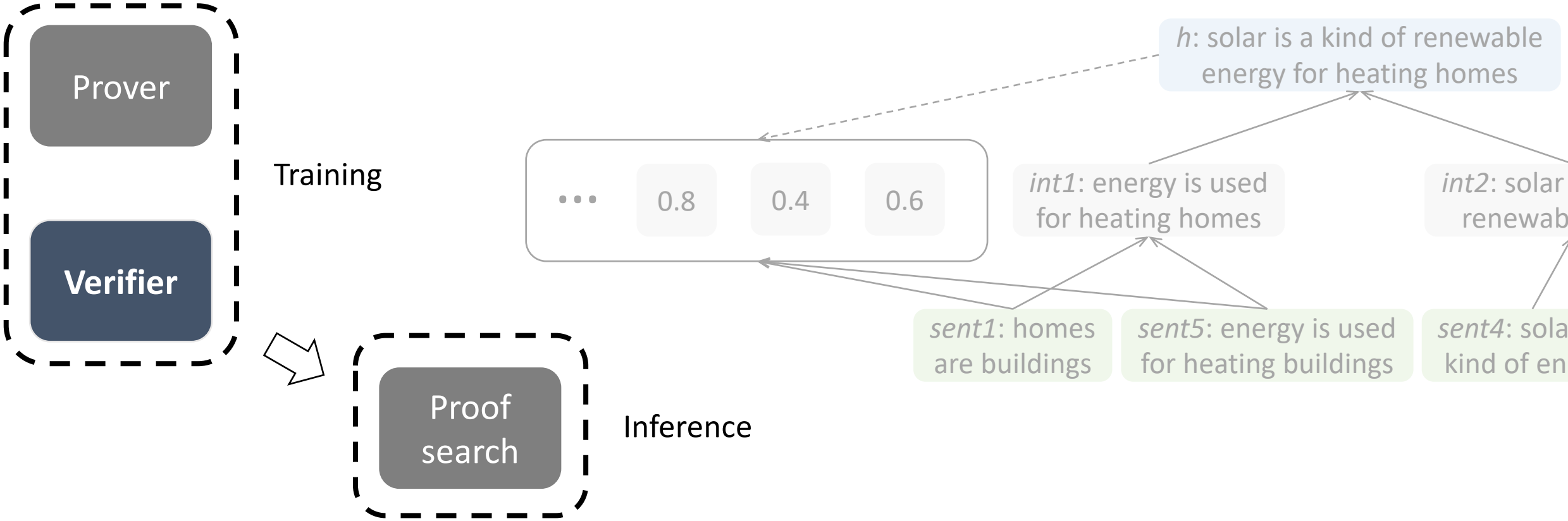
*sent1*: homes are buildings
*sent2*: solar is renewable
*sent3*: wind is a kind of energy
*sent4*: solar is a kind of energy
*sent5*: energy is used for heating buildings
*sent6*: coal is nonrenewable
…
…



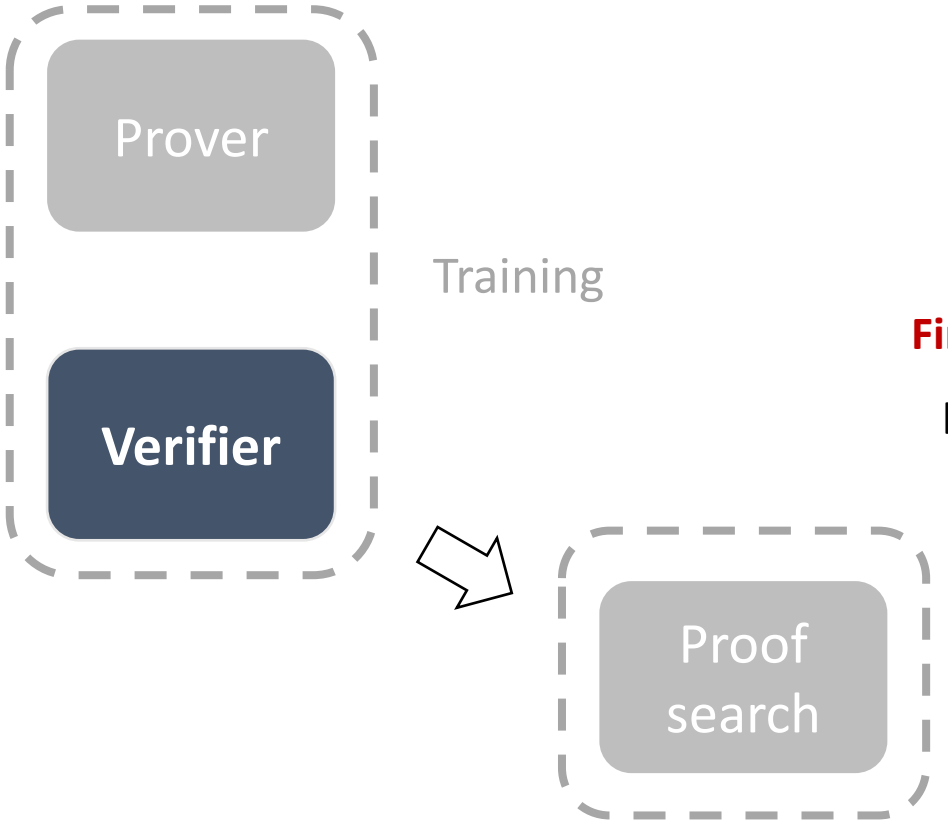
# NLProofS: Natural Language Proof Search

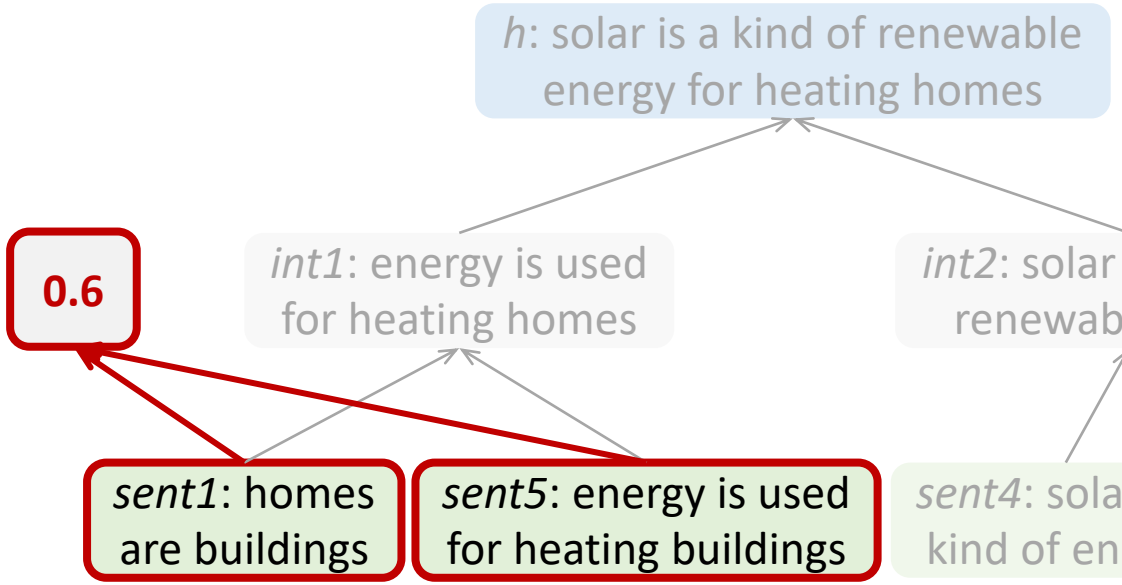


# NLProofS: Natural Language Proof Search

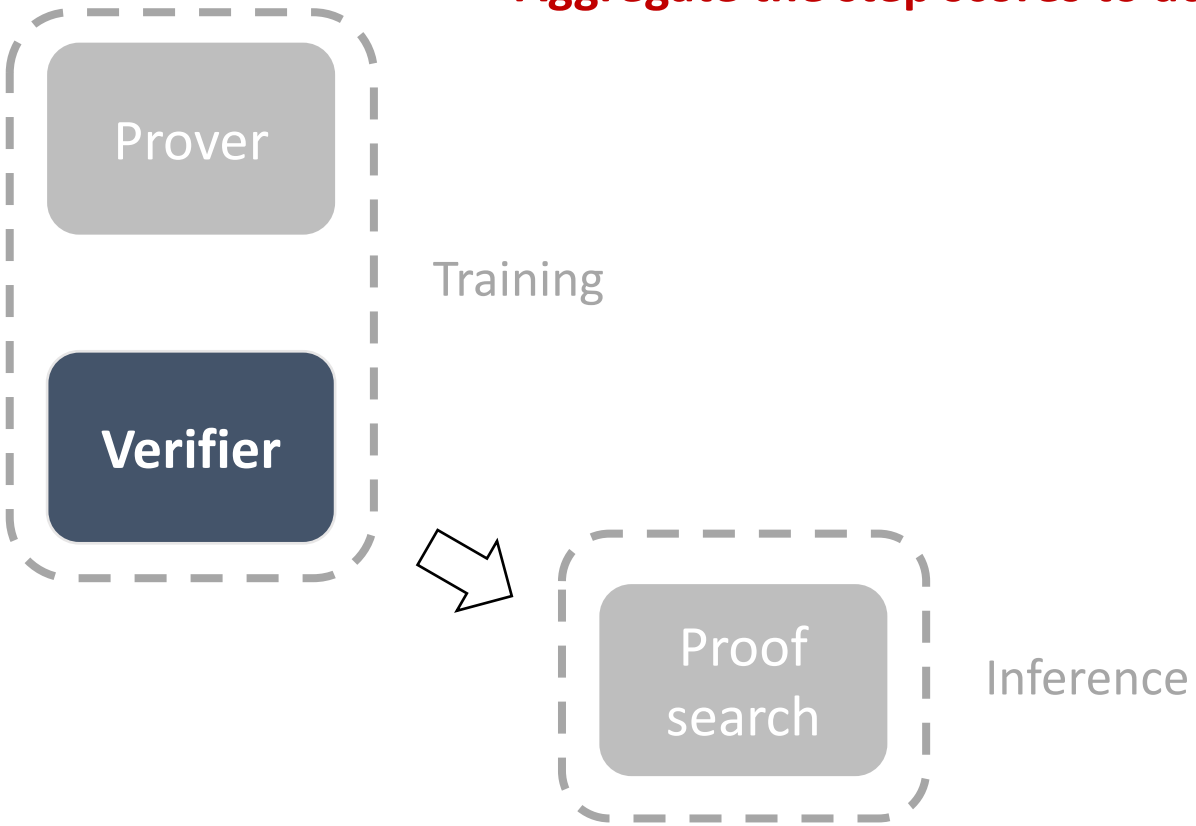


# NLProofS: Natural Language Proof Search

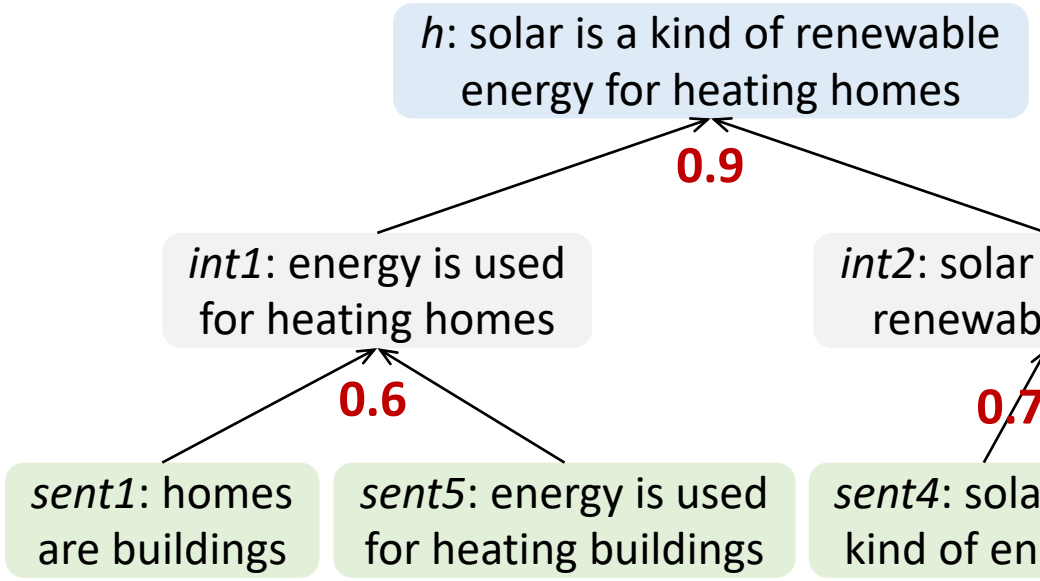


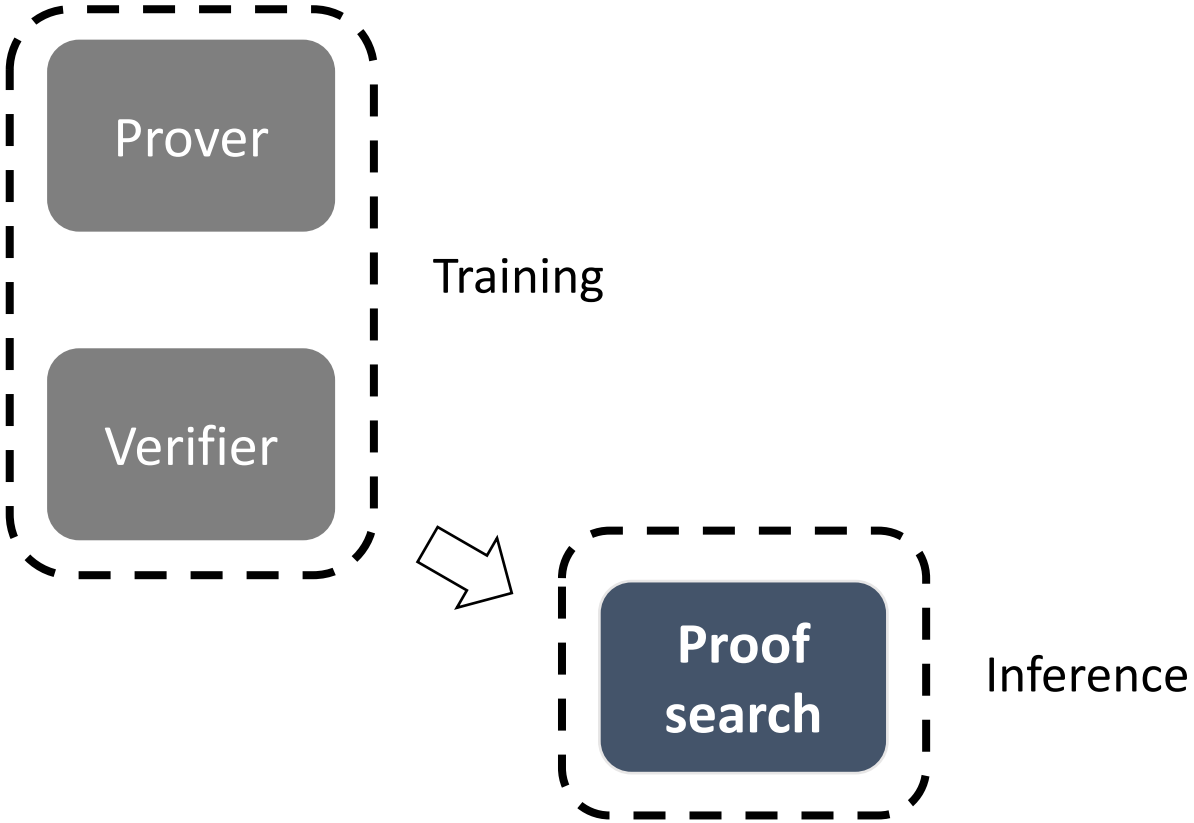
# NLProofS: Natural Language Proof Search

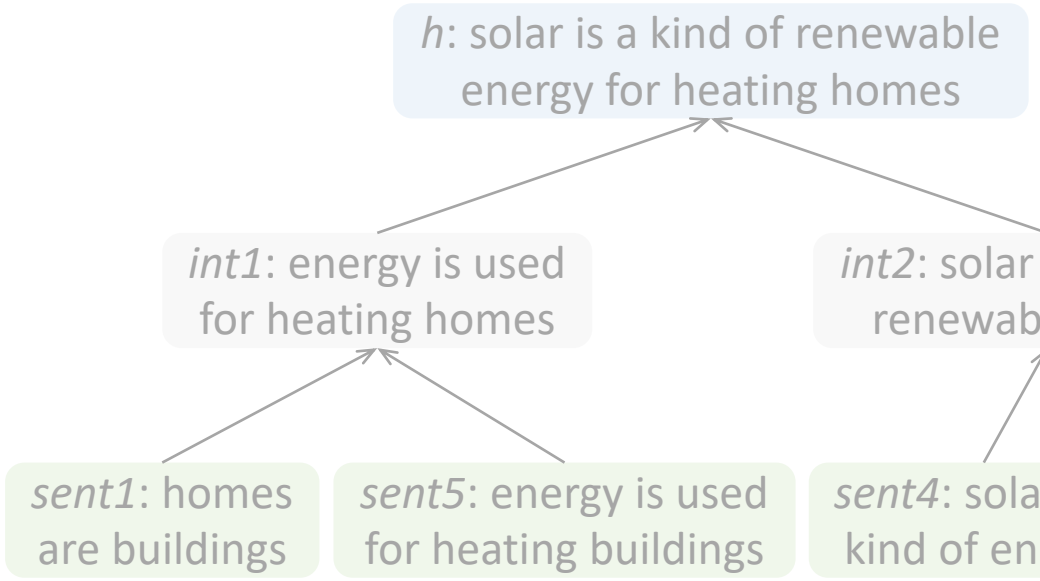



**Aggregate the step scores to across the entire proof**

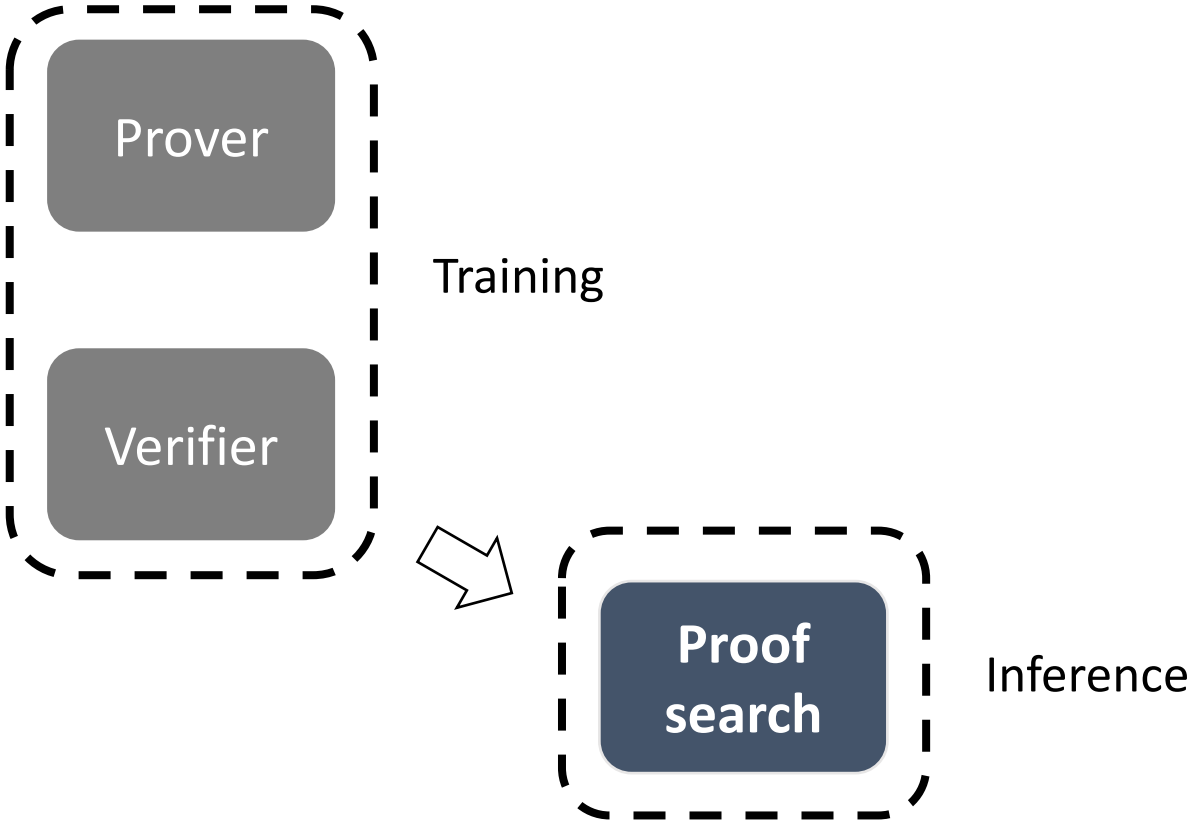


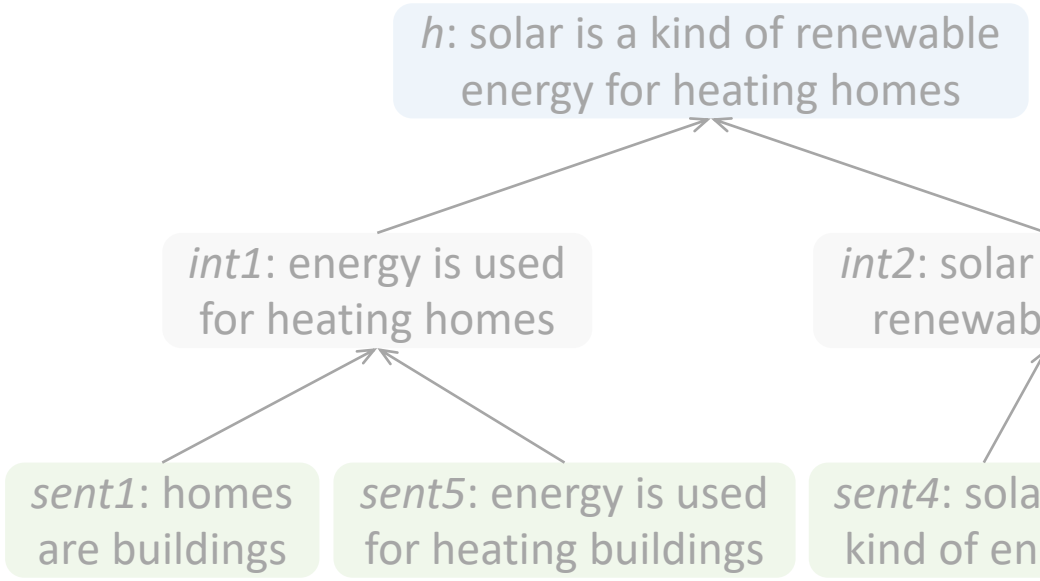
# NLProofS: Natural Language Proof Search



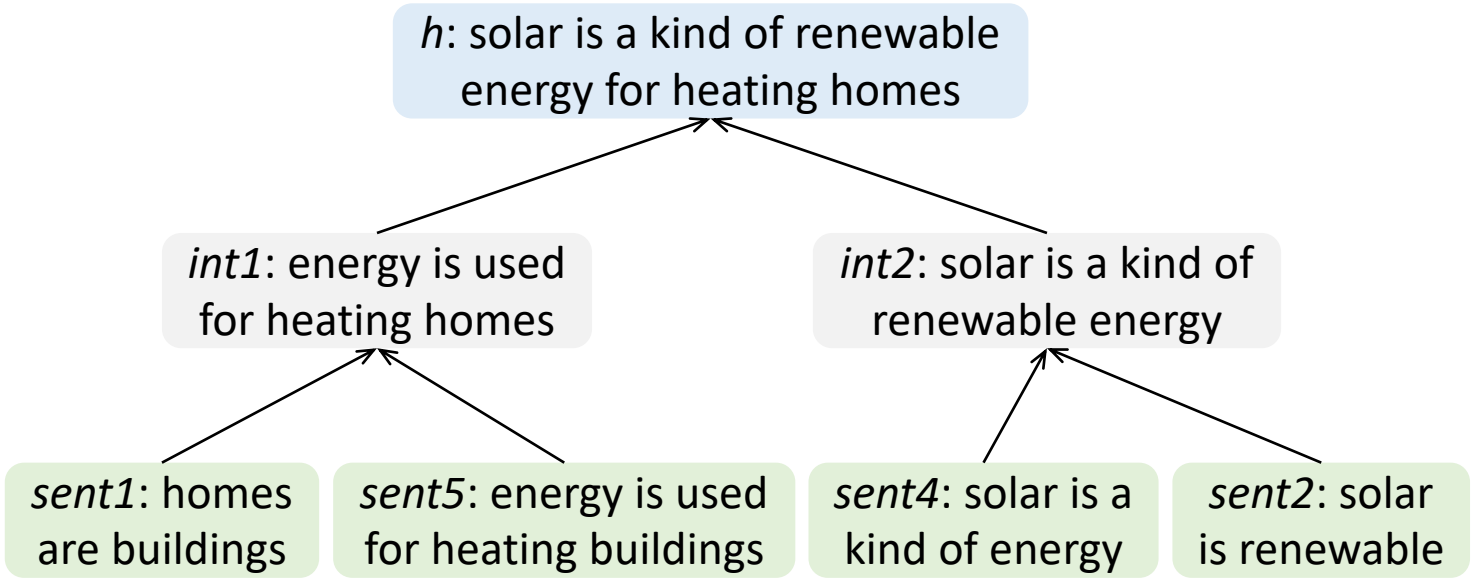
# NLProofS: Natural Language Proof Search



# NLProofS: Natural Language Proof Search

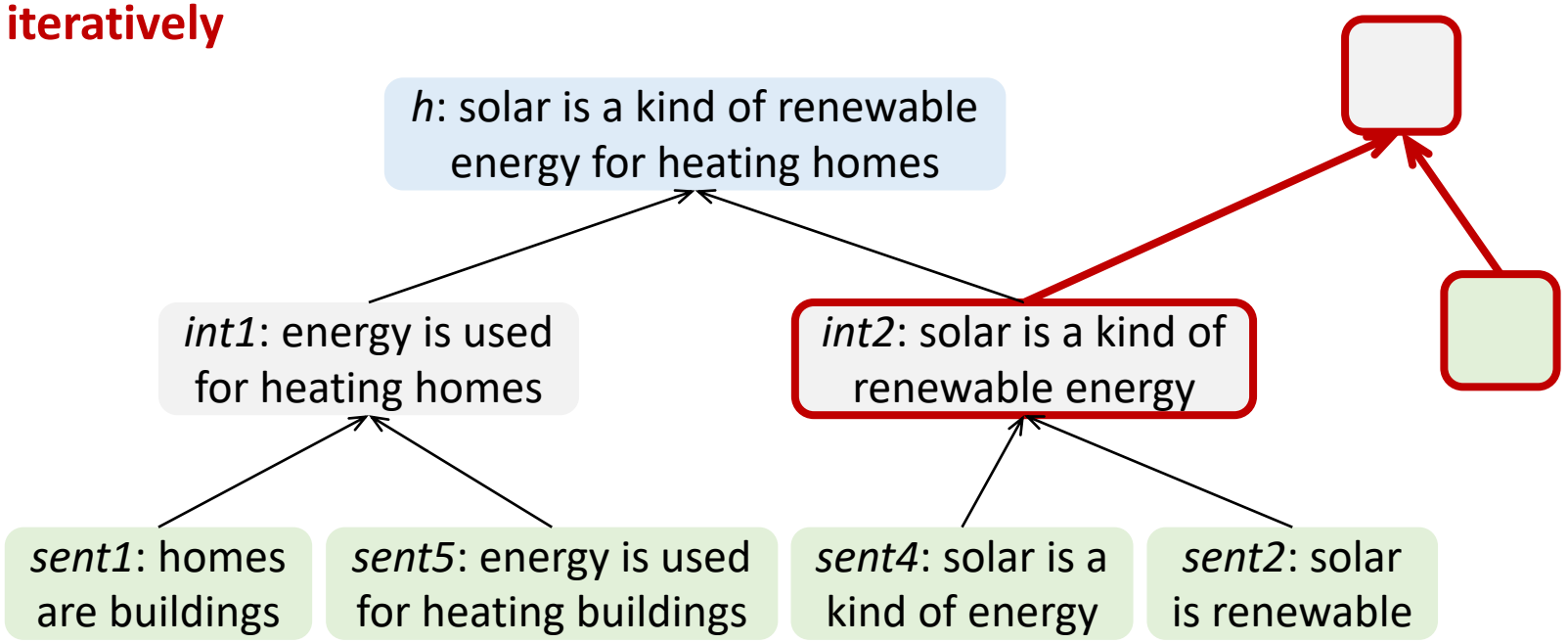
**1. Initialization: a proof generated by the prover alone**




**Proof search**

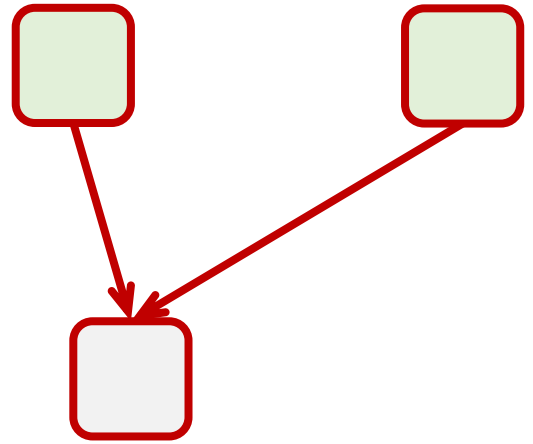
# NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. **Iteration: expand the graph iteratively**

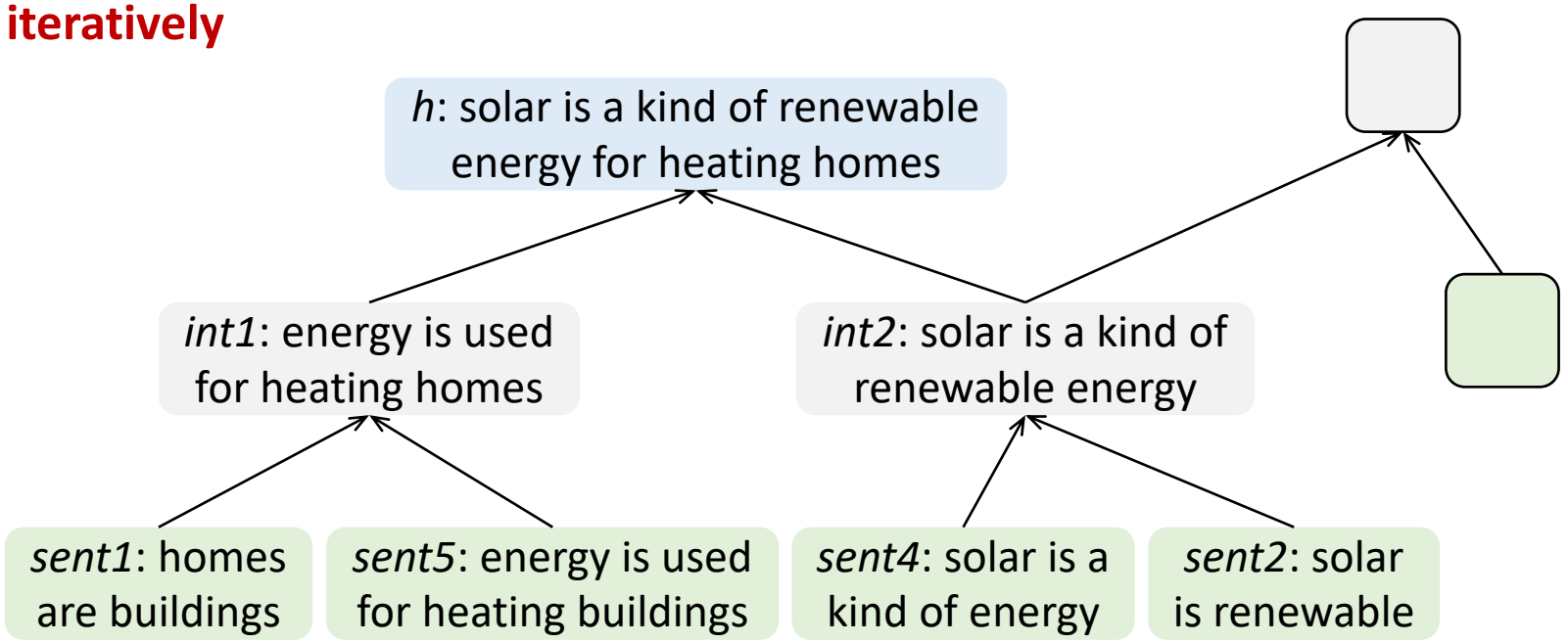



**Proof search**

# NLProofS: Natural Language Proof Search



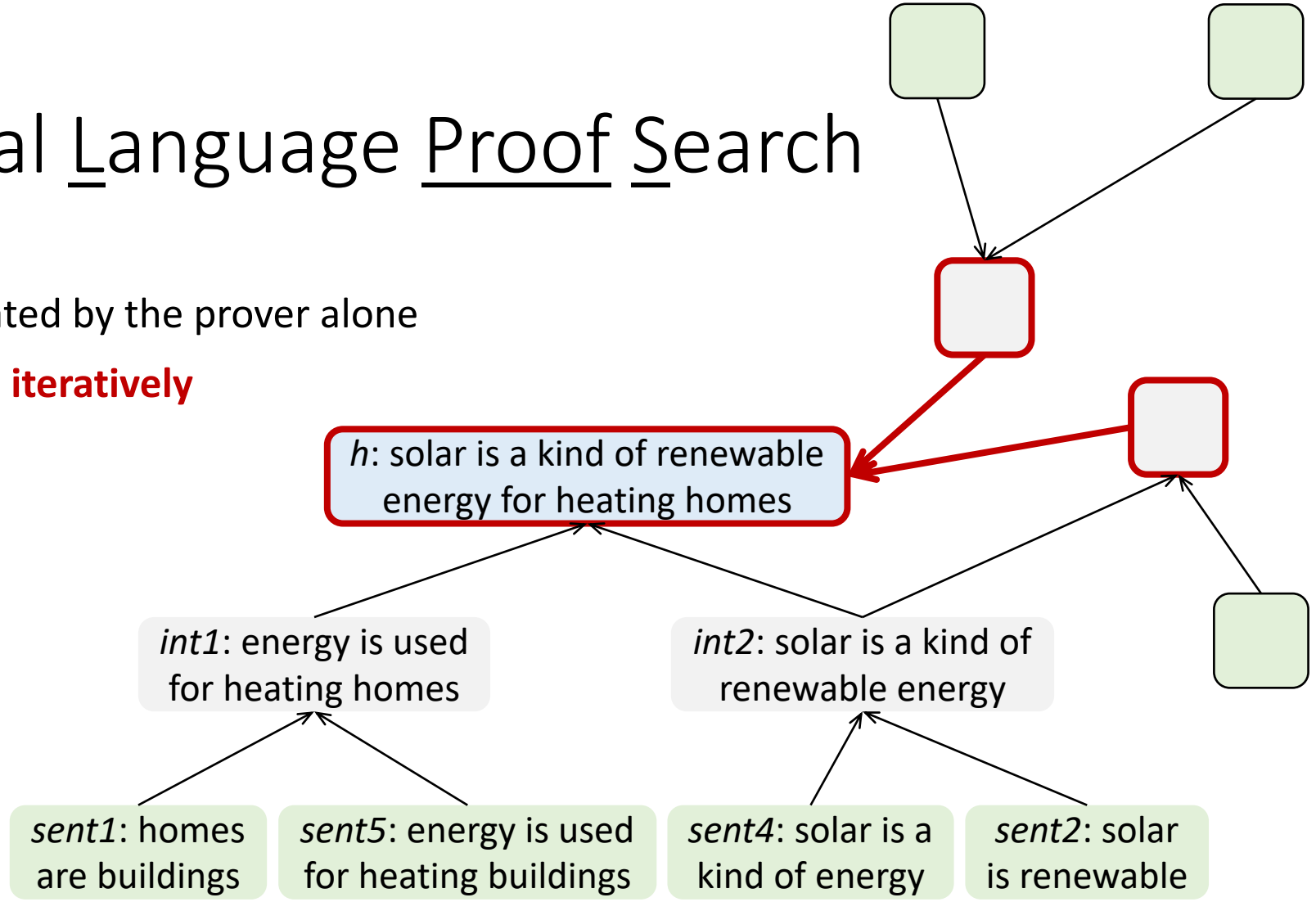
1. Initialization: a proof generated by the prover alone
2. **Iteration: expand the graph iteratively**




**Proof search**

# NLProofS: Natural Language Proof Search

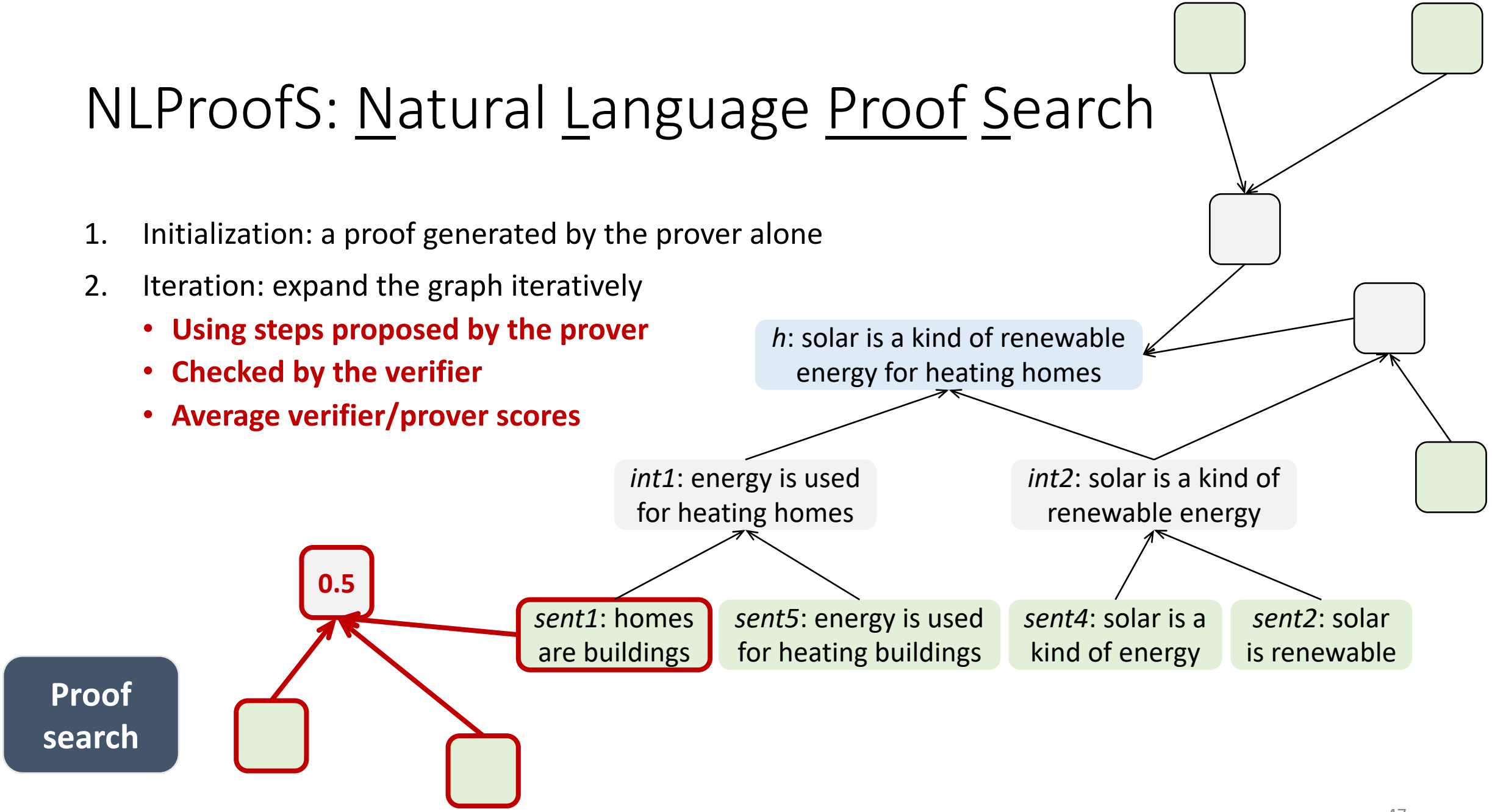
1. Initialization: a proof generated by the prover alone
2. **Iteration: expand the graph iteratively**




**Proof search**

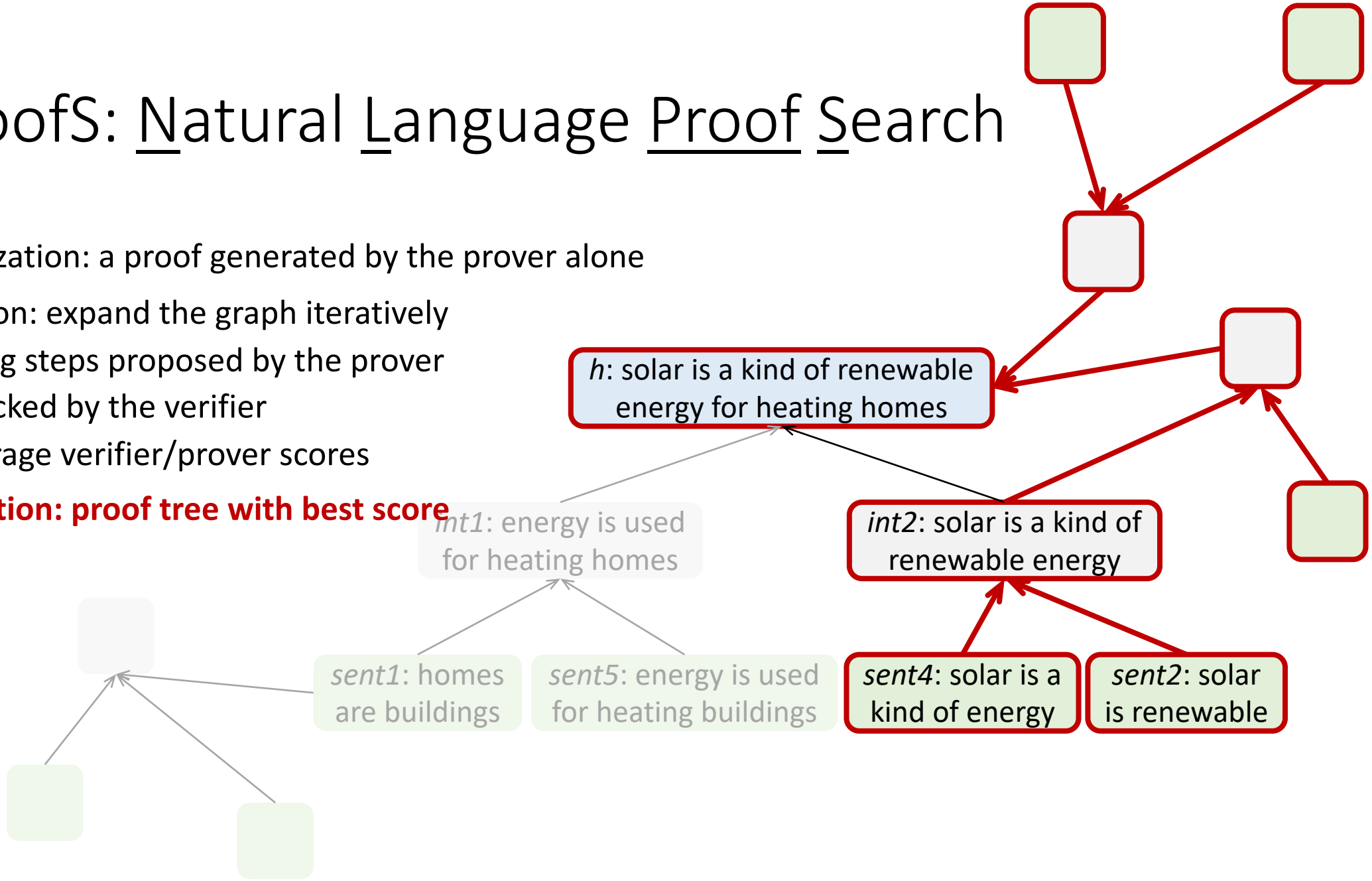
# NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively
   - **Using steps proposed by the prover**
   - **Checked by the verifier**
   - **Average verifier/prover scores**



# NLProofS: Natural Language Proof Search

1. Initialization: a proof generated by the prover alone
2. Iteration: expand the graph iteratively
   - Using steps proposed by the prover
   - Checked by the verifier
   - Average verifier/prover scores
3. **Extraction: proof tree with best score**




**Proof search**

# Experiments

- Evaluate on two benchmarks
  - **RuleTaker**: Simple, synthetic proofs [Tafjord et al. Findings of ACL 2021]
  - **EntailmentBank**: ~2K challenging, human-written proofs [Dalvi et al. EMNLP 2021]
- State-of-the-art results on both
- Ablations highlight the importance of the verifier

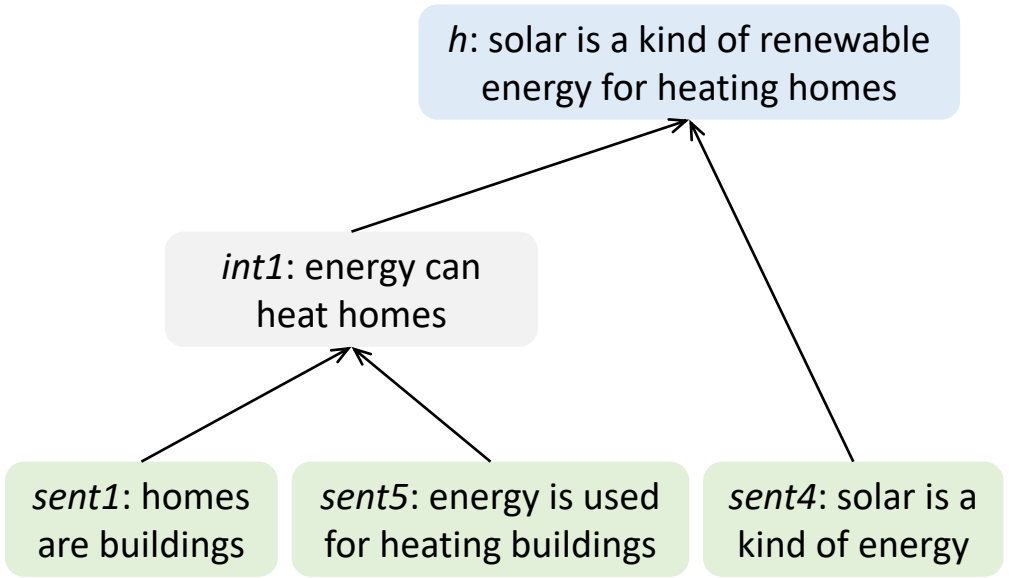
# Experiments

- Evaluate on two benchmarks
  - **RuleTaker**: Simple, synthetic proofs [Tafjord et al. Findings of ACL 2021]
  - **EntailmentBank**: ~2K challenging, human-written proofs [Dalvi et al. EMNLP 2021]
- State-of-the-art results on both
- Ablations highlight the importance of the verifier

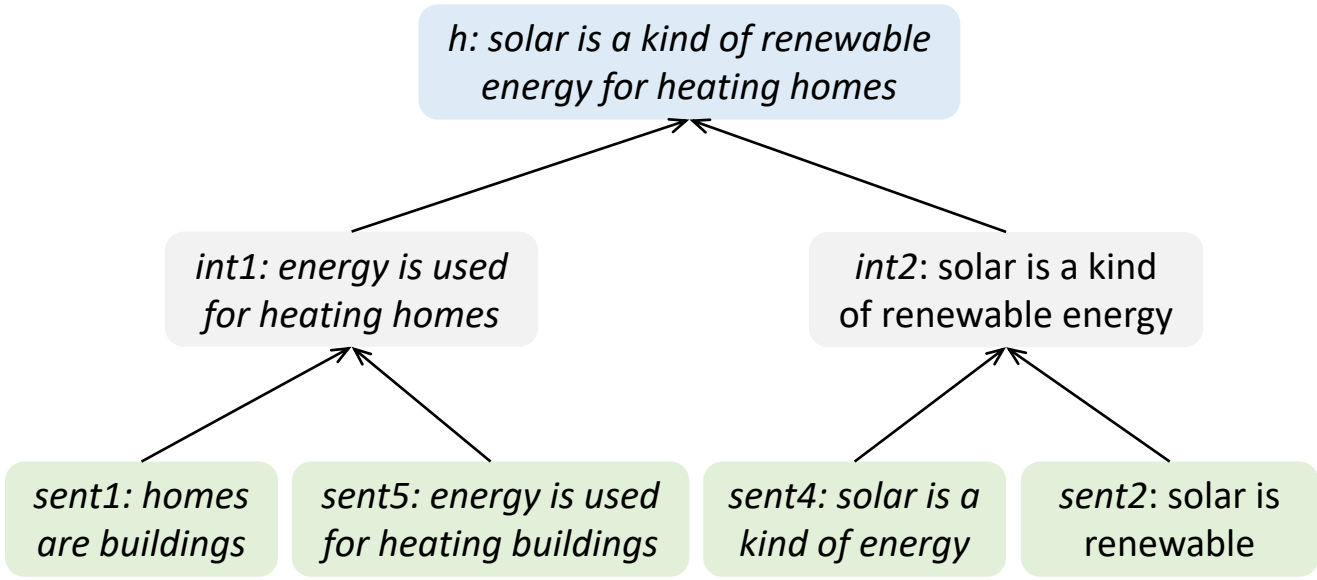
**25 supporting facts, including distractors**

[Dalvi et al. EMNLP 2021]

# EntailmentBank: Evaluation Metrics




Predicted proof tree

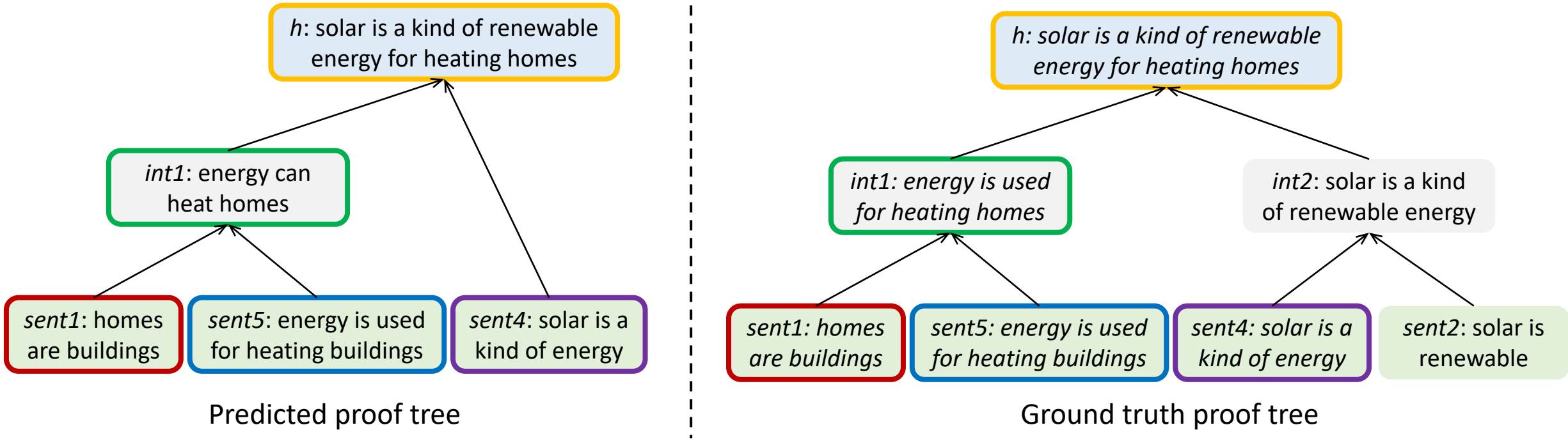



Ground truth proof tree

- EntailmentBank's four official metrics: Leaves, Steps, Intermediates, Overall

[Dalvi et al. EMNLP 2021]

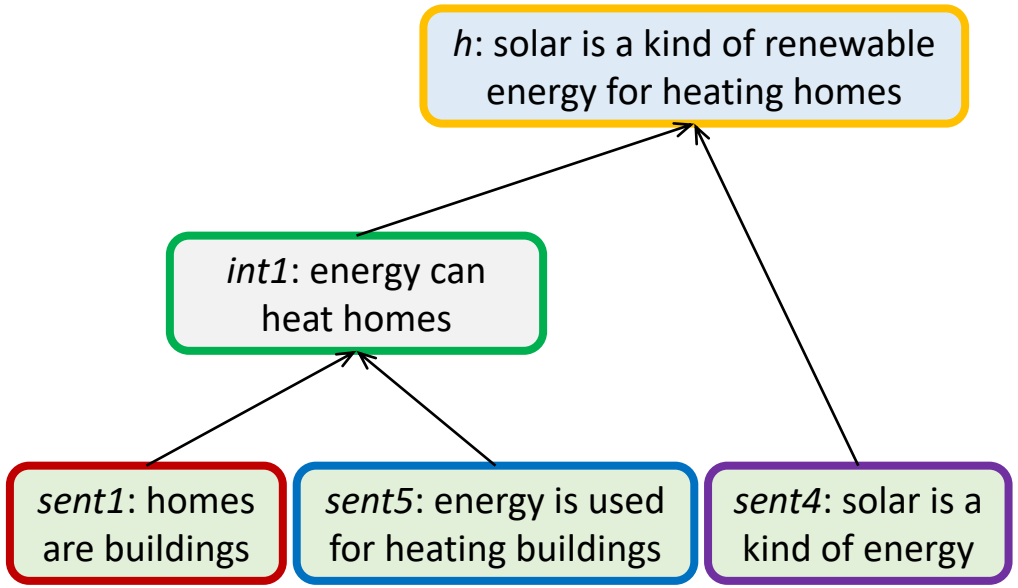
# EntailmentBank: Evaluation Metrics



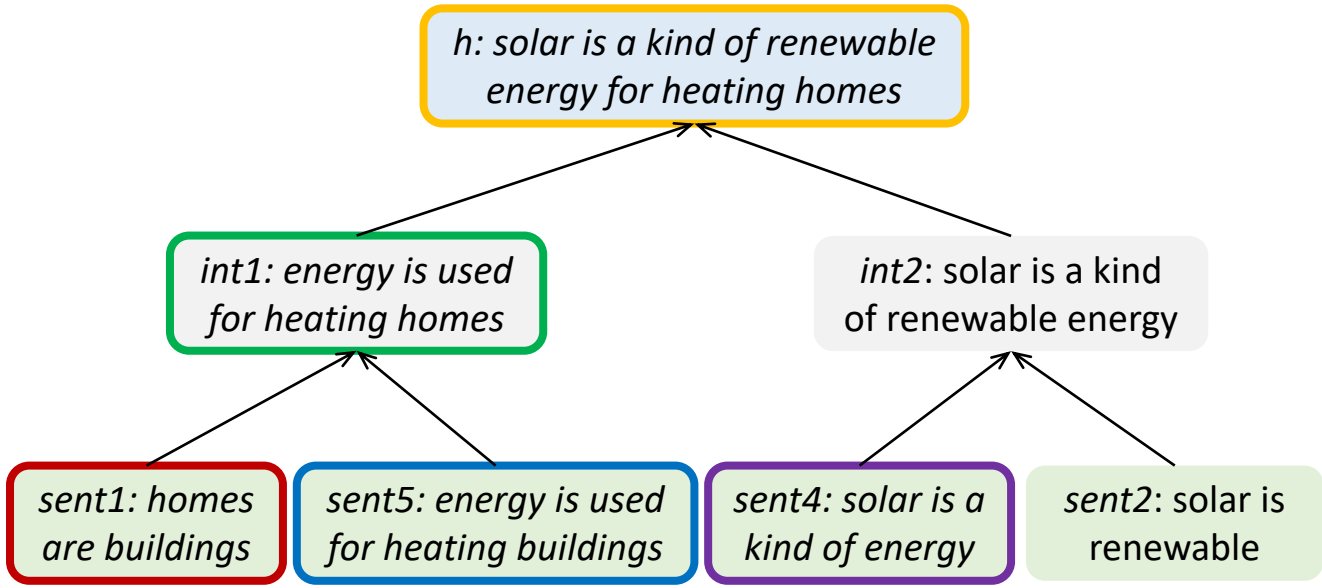

- EntailmentBank's four official metrics: Leaves, Steps, Intermediates, Overall
- **Based on heuristic matching between the nodes**

[Dalvi et al. EMNLP 2021]

# EntailmentBank: Evaluation Metrics



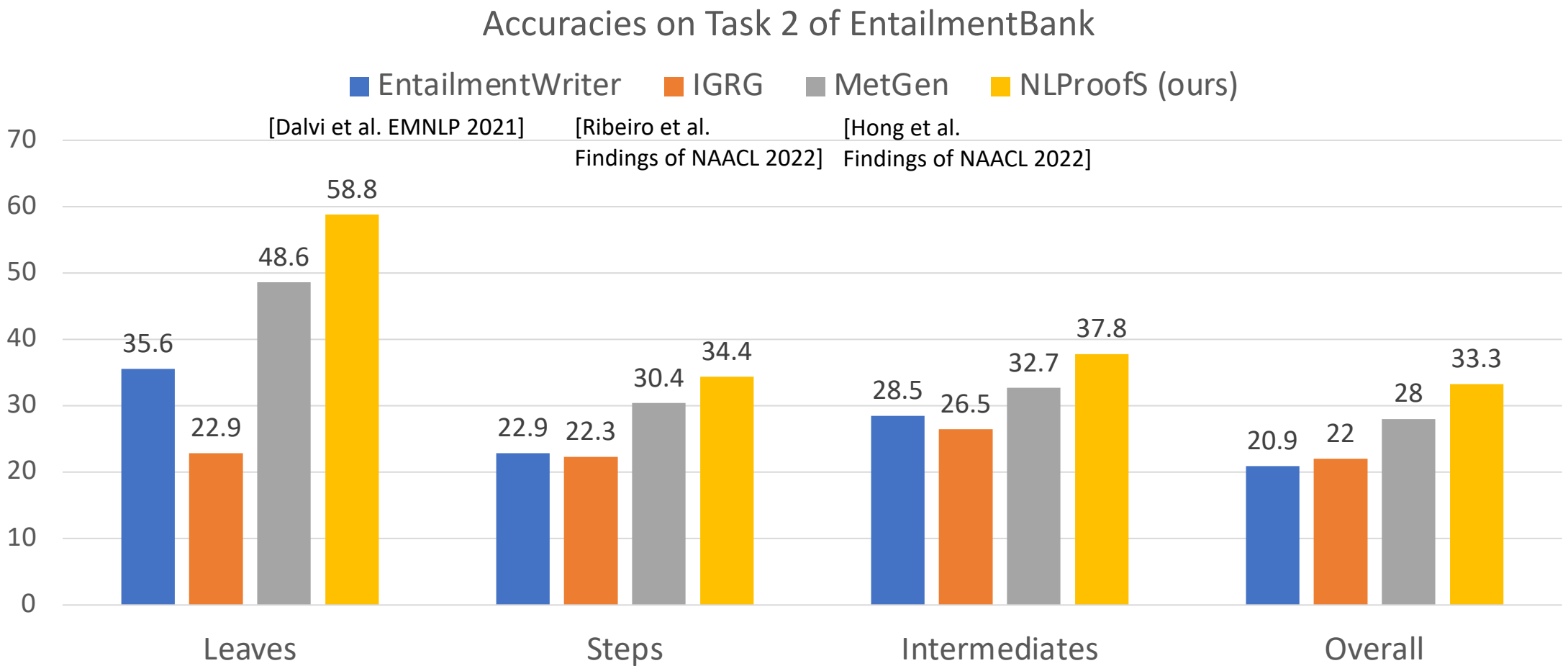

Predicted proof tree



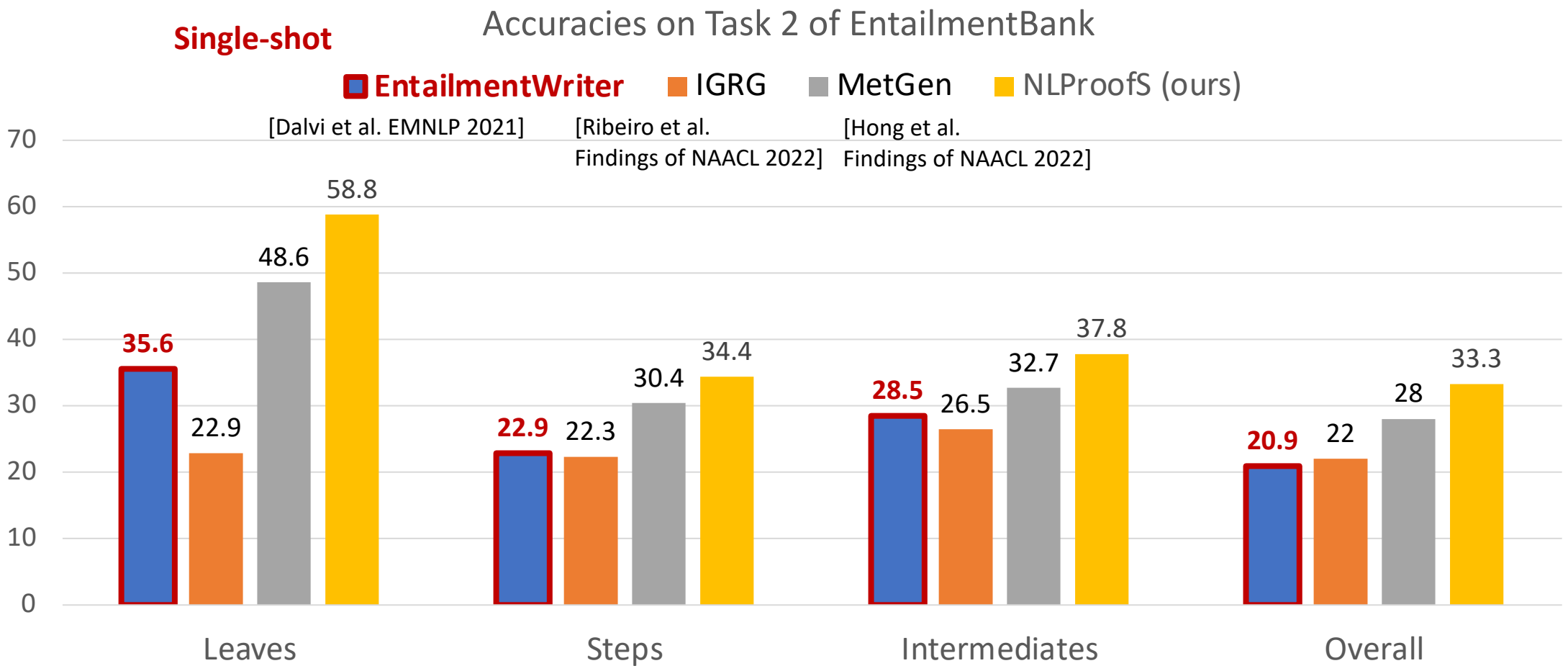

Ground truth proof tree

- EntailmentBank's four official metrics: Leaves, Steps, Intermediates, Overall
- Based on heuristic matching between the nodes
- **Limitations: Cannot handle correct predictions different from the ground truth**

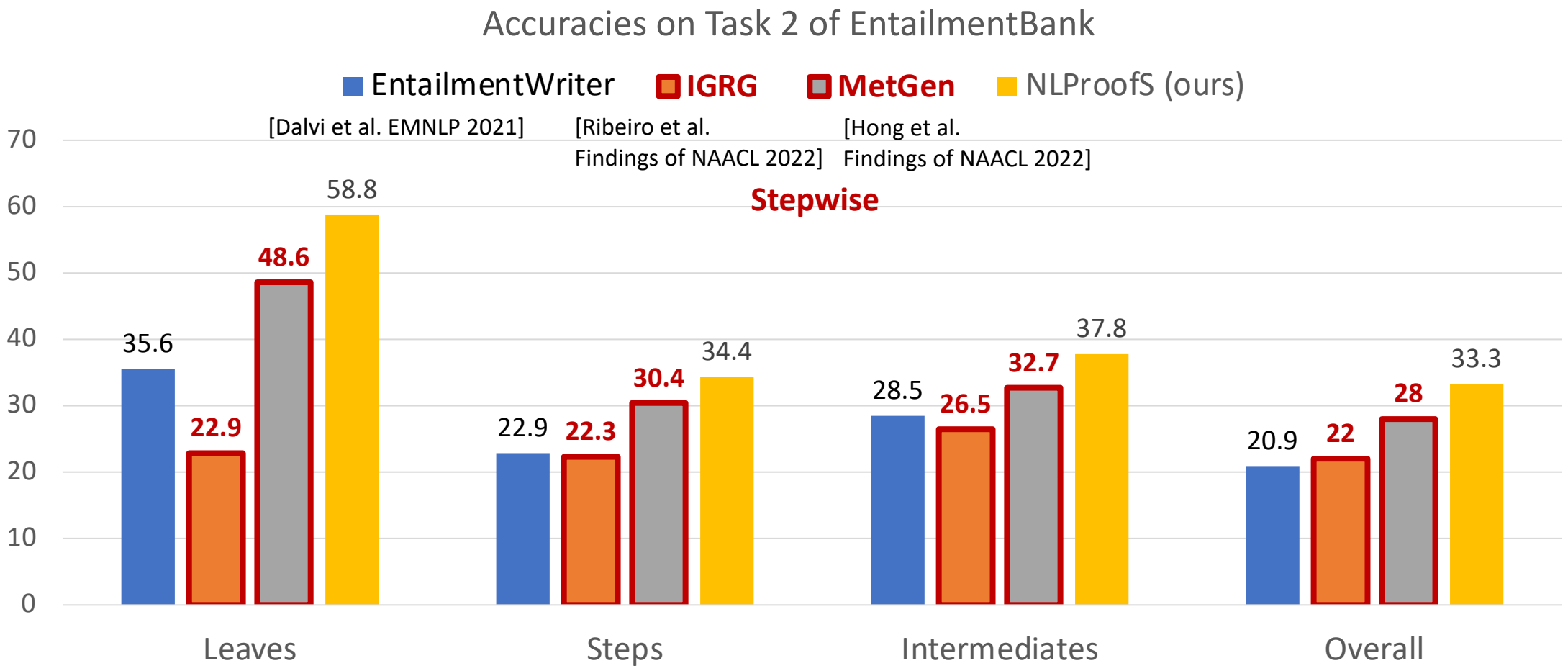
# State-of-the-art Performance on EntailmentBank



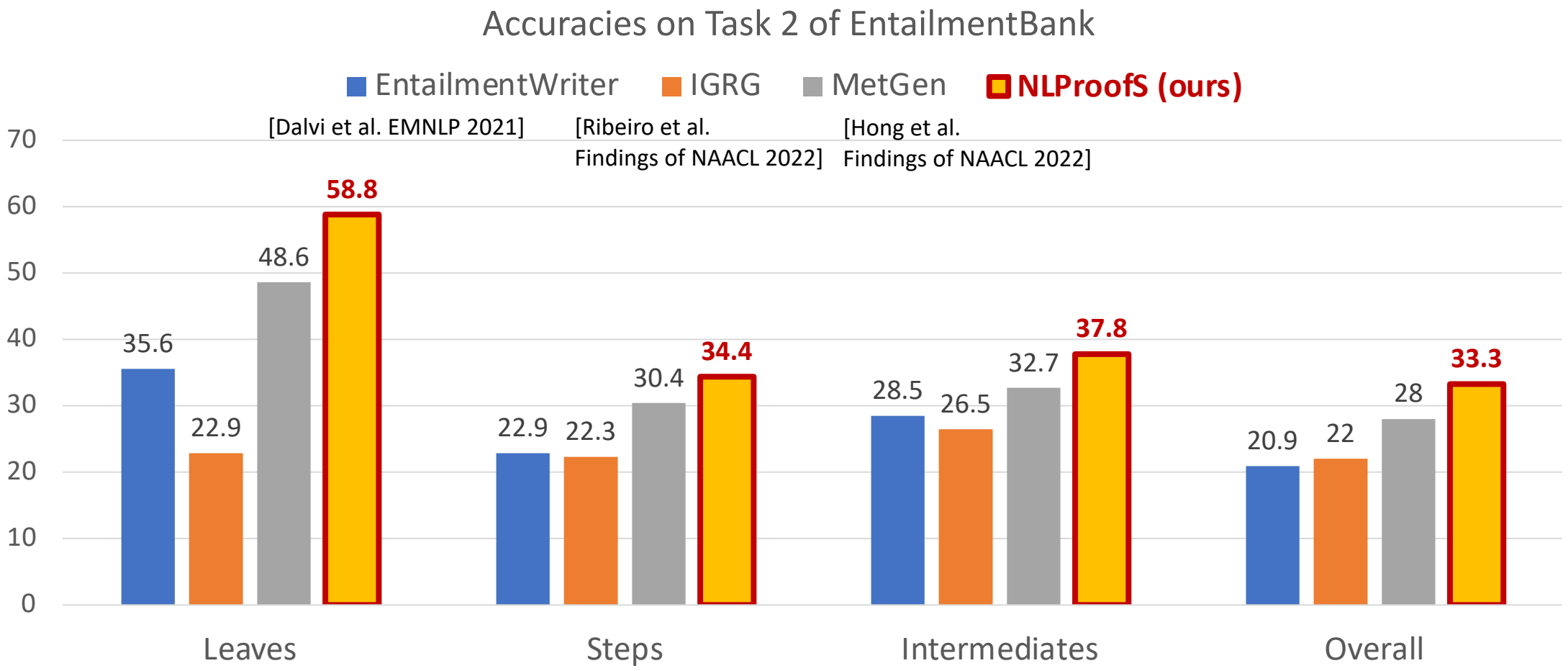
# State-of-the-art Performance on EntailmentBank



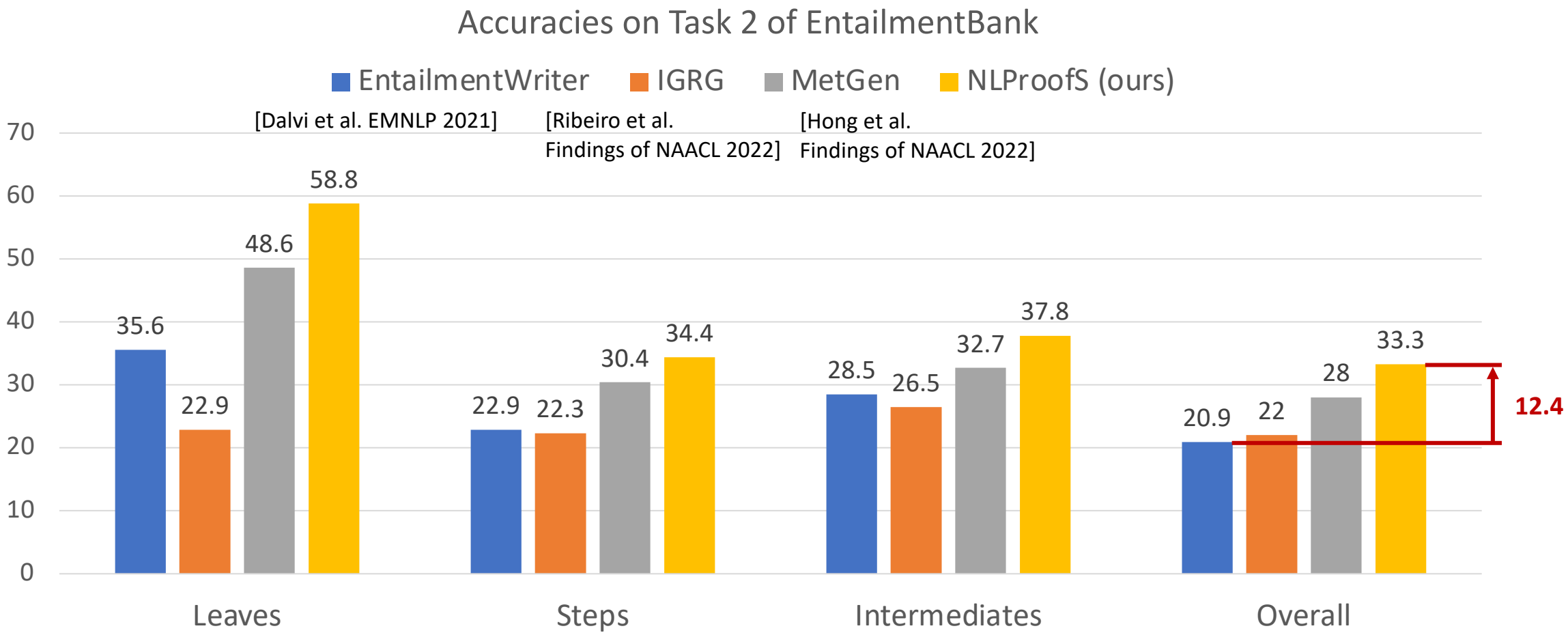
# State-of-the-art Performance on EntailmentBank



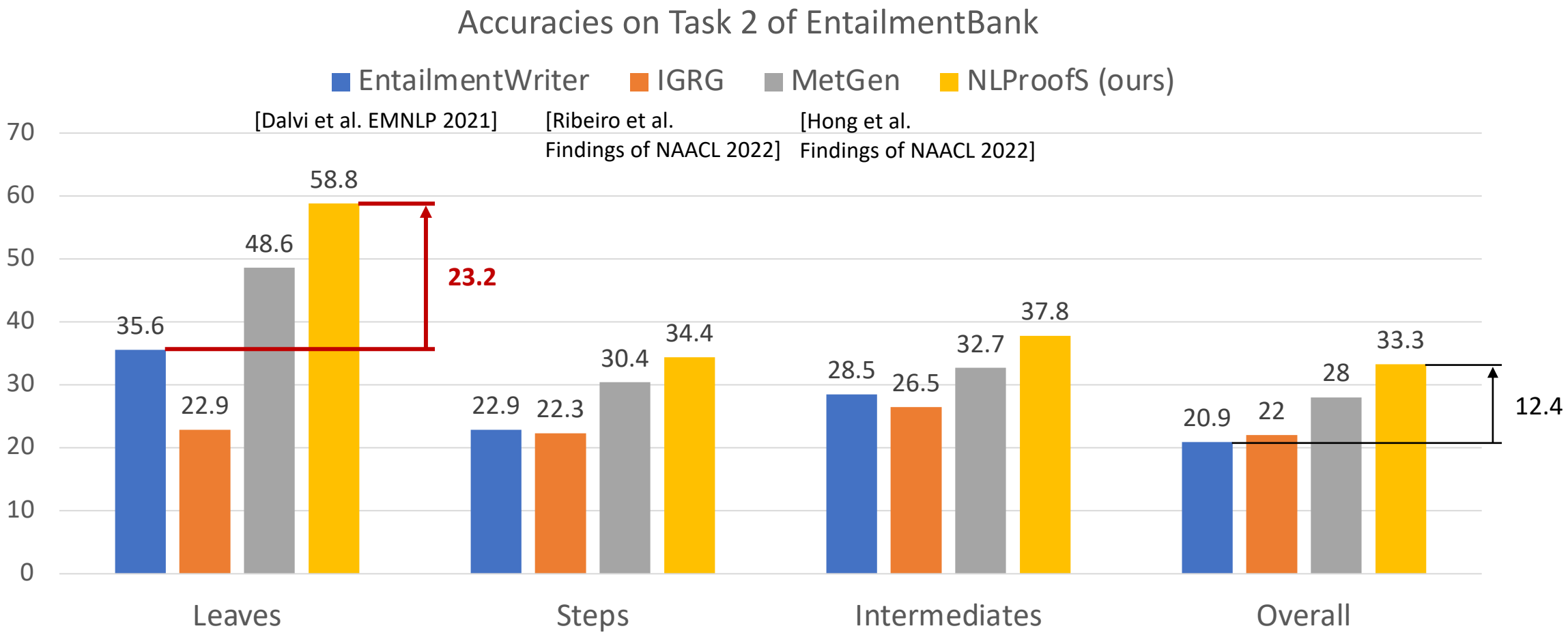
# State-of-the-art Performance on EntailmentBank



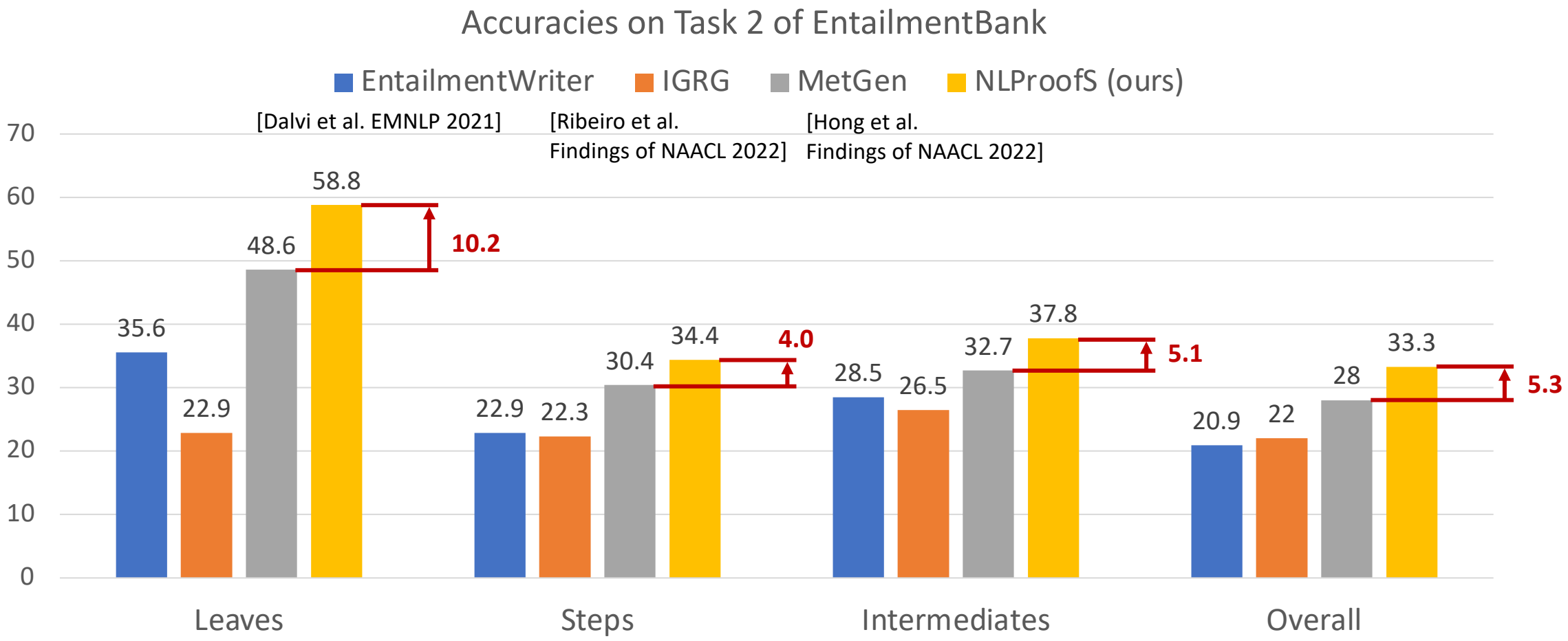
# State-of-the-art Performance on EntailmentBank



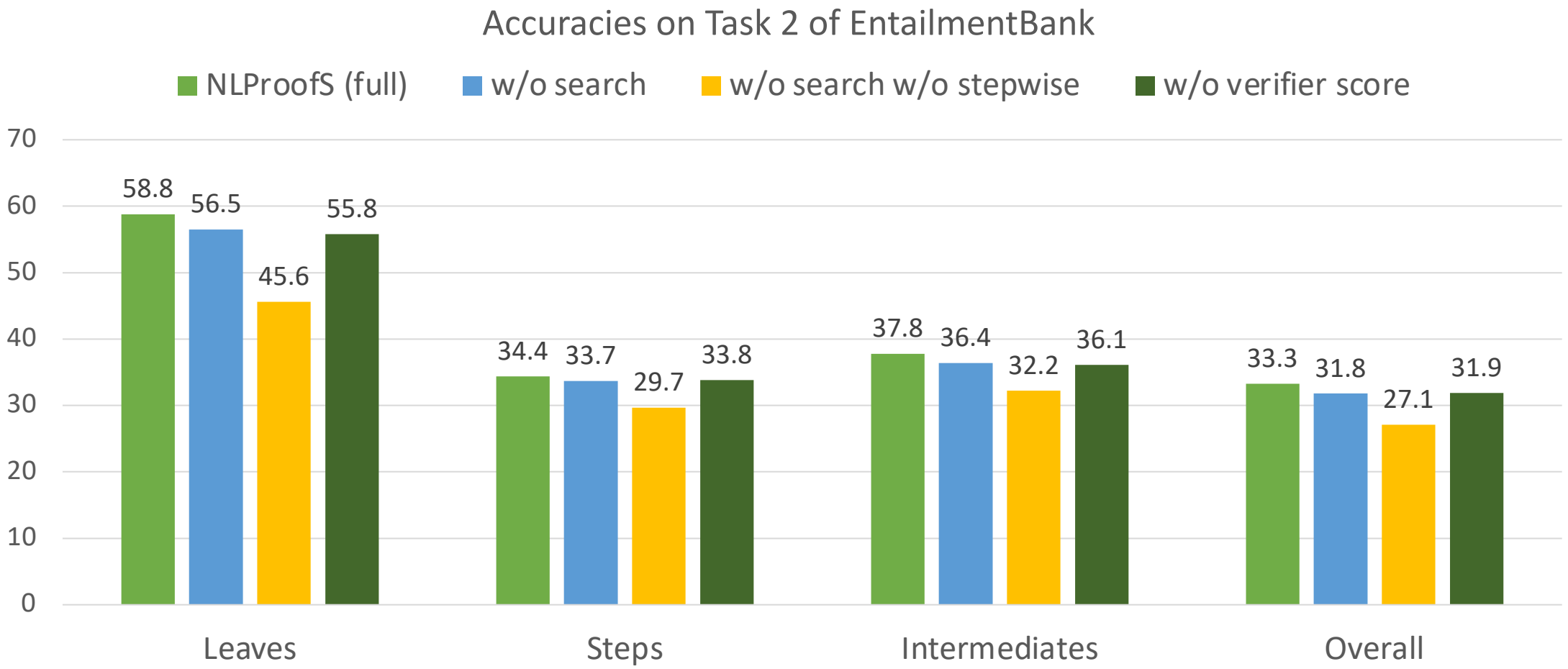
# State-of-the-art Performance on EntailmentBank



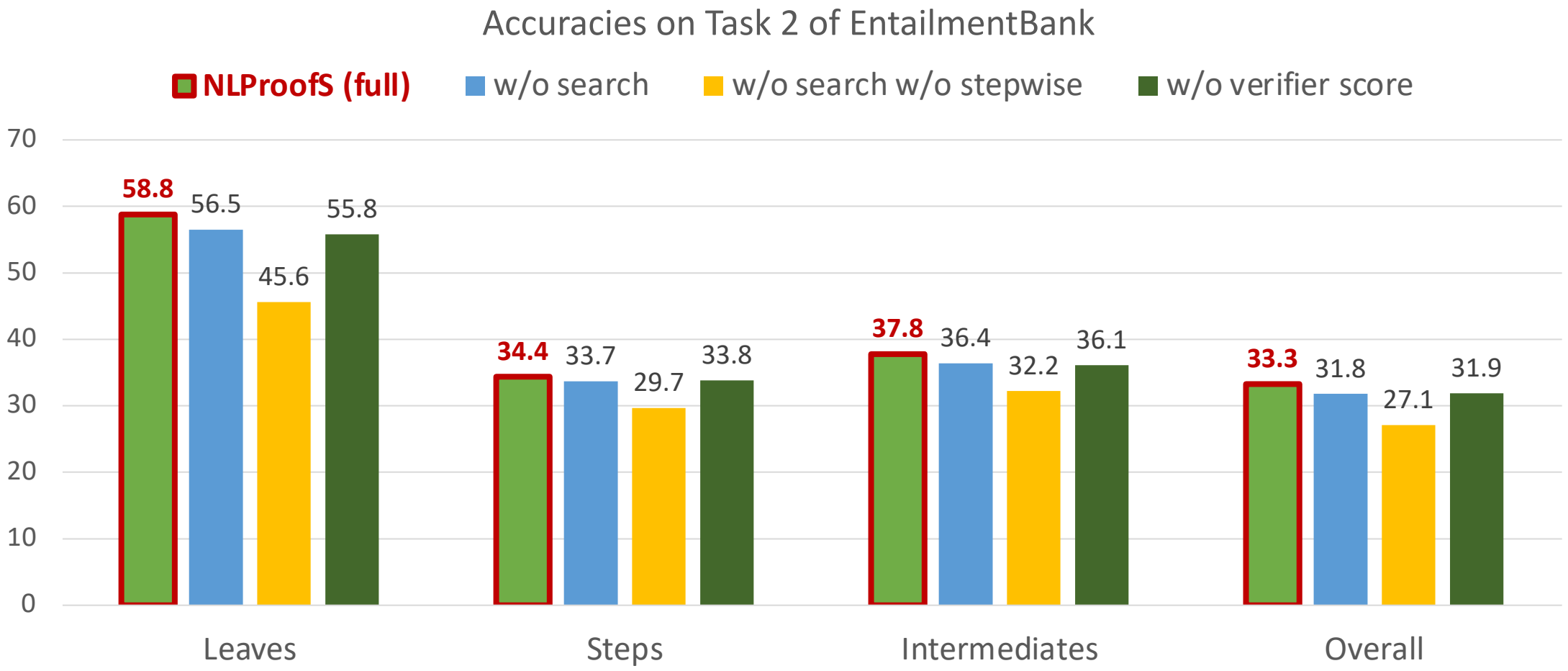
# State-of-the-art Performance on EntailmentBank



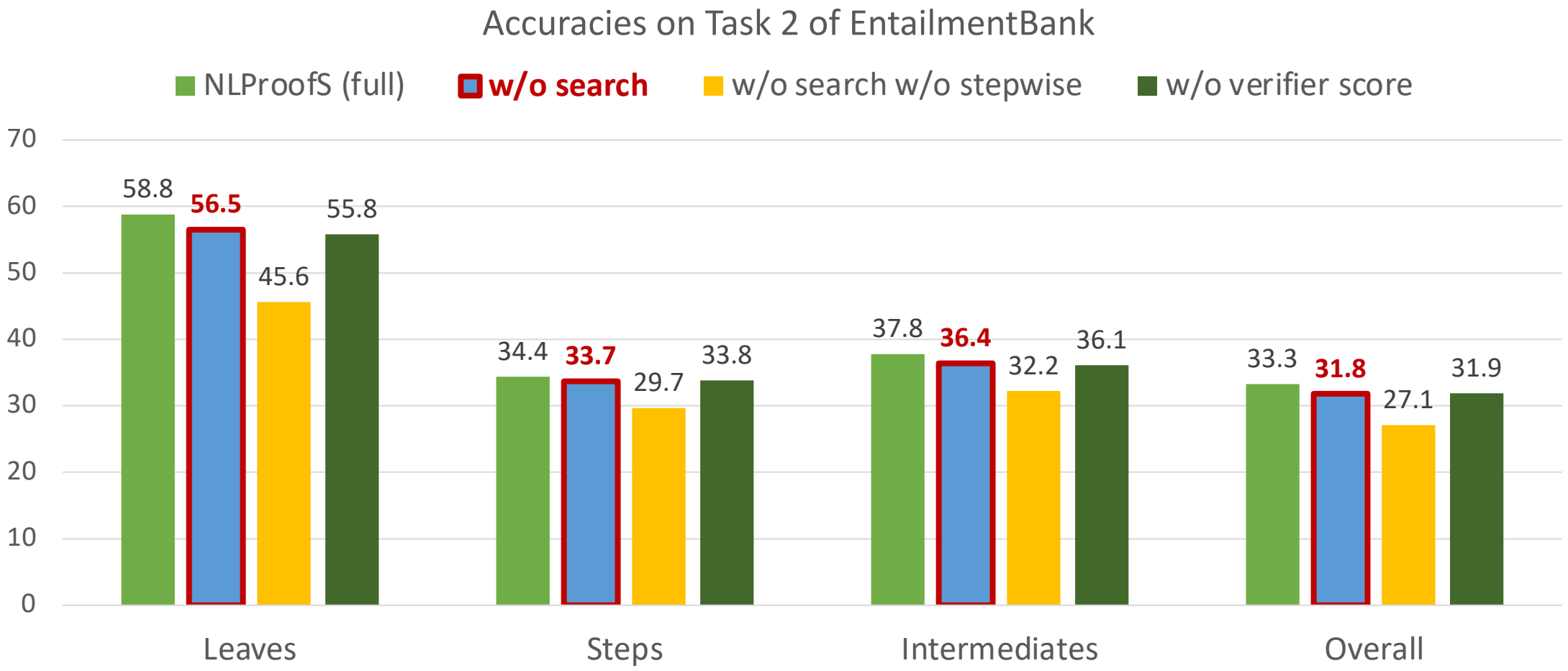
# Ablations


Accuracies on Task 2 of EntailmentBank
NLProofS (full)
w/o search
w/o search w/o stepwise
w/o verifier score
70
60
50
40
30
20
10
0
58.8
56.5
45.6
55.8
34.4
33.7
29.7
33.8
37.8
36.4
32.2
36.1
33.3
31.8
27.1
31.9
Leaves
Steps
Intermediates
Overall

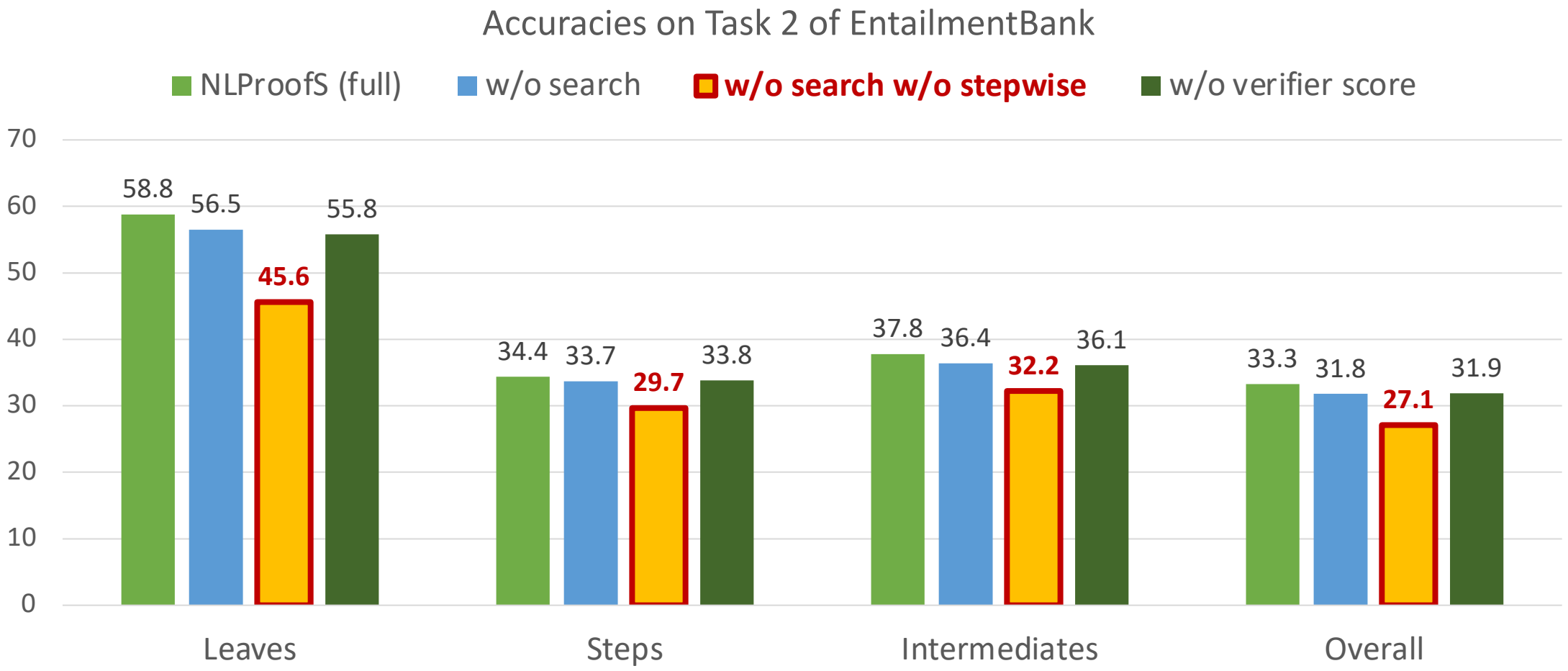
# Ablations


Accuracies on Task 2 of EntailmentBank
NLProofS (full)
w/o search
w/o search w/o stepwise
w/o verifier score
70
60
50
40
30
20
10
0
58.8
56.5
45.6
55.8
34.4
33.7
29.7
33.8
37.8
36.4
32.2
36.1
33.3
31.8
27.1
31.9
Leaves
Steps
Intermediates
Overall

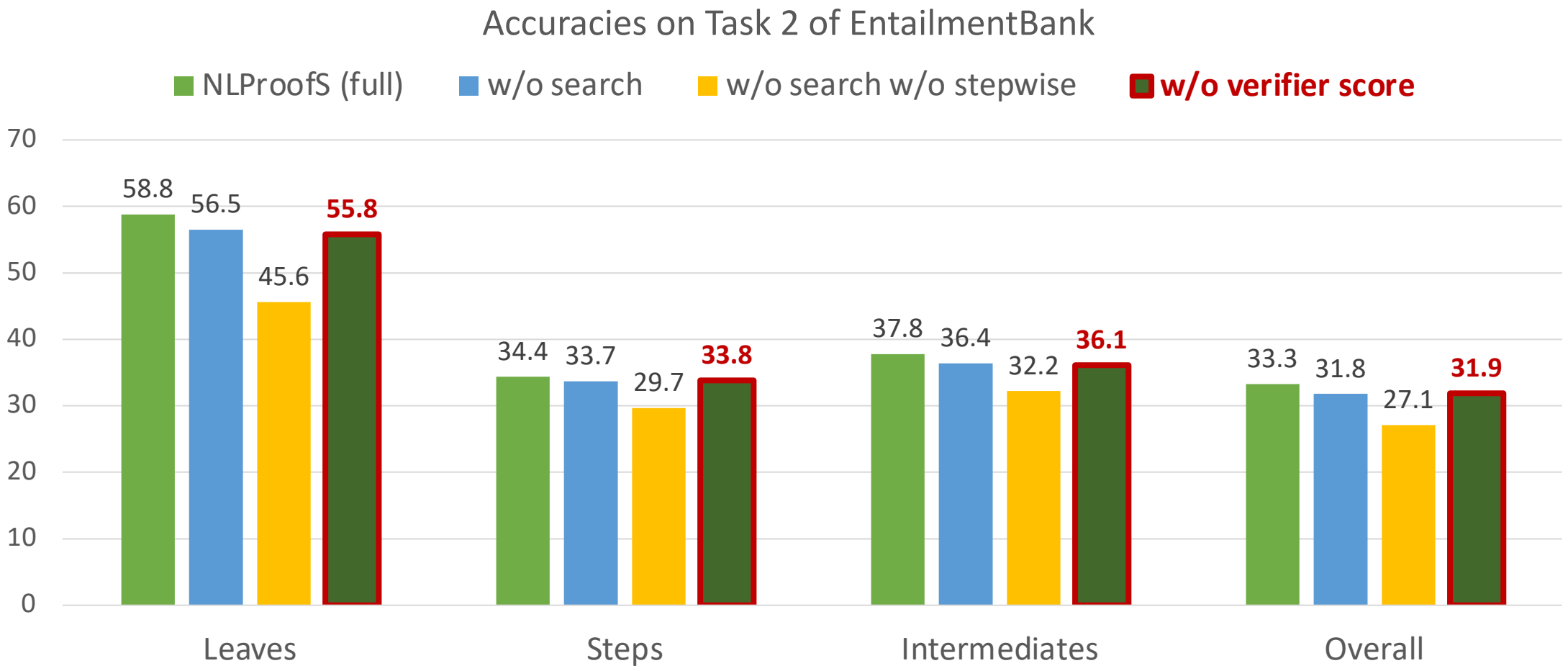
# Verifier-Guided Proof Search Is Helpful



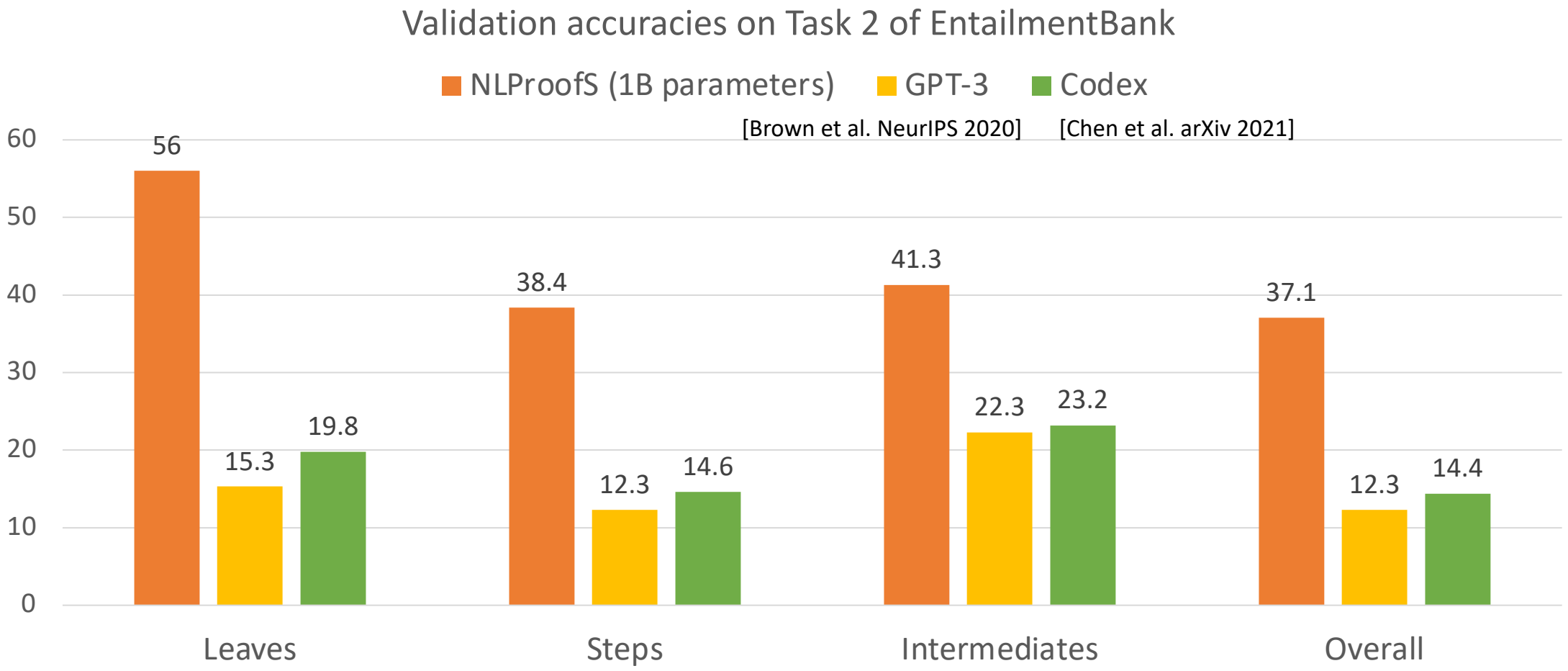
# Stepwise Generation Is Helpful



# The Verifier Is Necessary for Proof Search



# Large Language Models w/ In-Context Learning



# Key Takeaways

- The verifier is important
  - **Prevent hallucinated generations**

# Key Takeaways

- The verifier is important
  - Prevent hallucinated generations
  - Also explored in other contexts, e.g., math word problems, code generation

[Cobbe et al. arXiv 2021] [Le and Wang et al. NeurIPS 2022]

# Generating Natural Language Proofs with Verifier-Guided Search

Kaiyu Yang, Jia Deng, Danqi Chen

PRINCETON UNIVERSITY



https://github.com/princeton-nlp/NLProofS