NLP Reading Group series: Towards Teachable Reasoning Systems

NLP Reading Group, University of Arizona

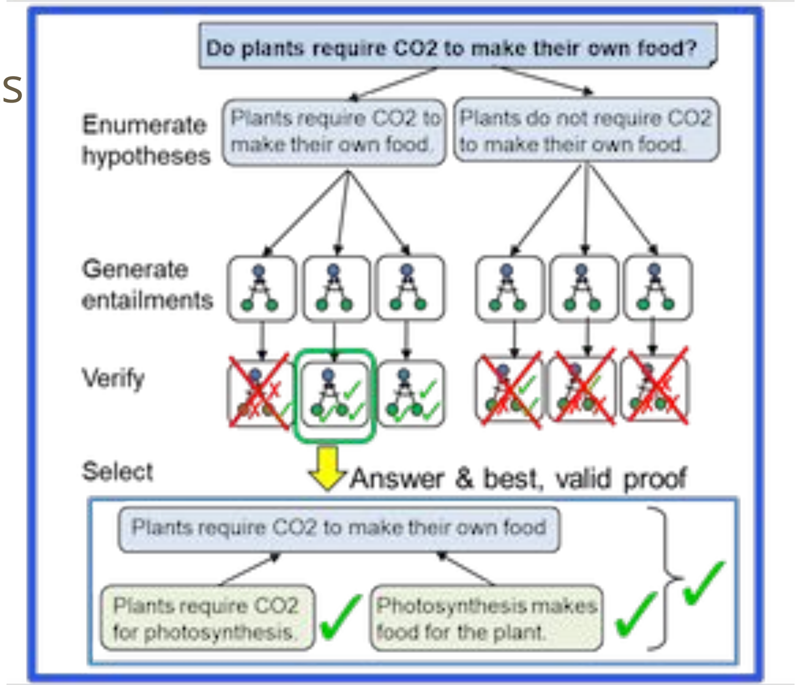
Sushma Akoju
Advisor: Prof. Mihai Surdeanu

NLP reading Group series: Towards Teachable Reasoning Systems

Towards Teachable Reasoning Systems : EMNLP 2021

1. Generates chains of reasoning
2. User correction with the beliefs & facts
3. Dynamic memory of corrected facts

https://arxiv.org/pdf/2204.13074.pdf


Do plants require CO2 to make their own food?
Enumerate hypotheses
Plants require CO2 to make their own food.
Plants do not require CO2 to make their own food.
Generate entailments
Verify
Select
Answer & best, valid proof
Plants require CO2 to make their own food
Plants require CO2 for photosynthesis.
Photosynthesis makes food for the plant.

Research Questions

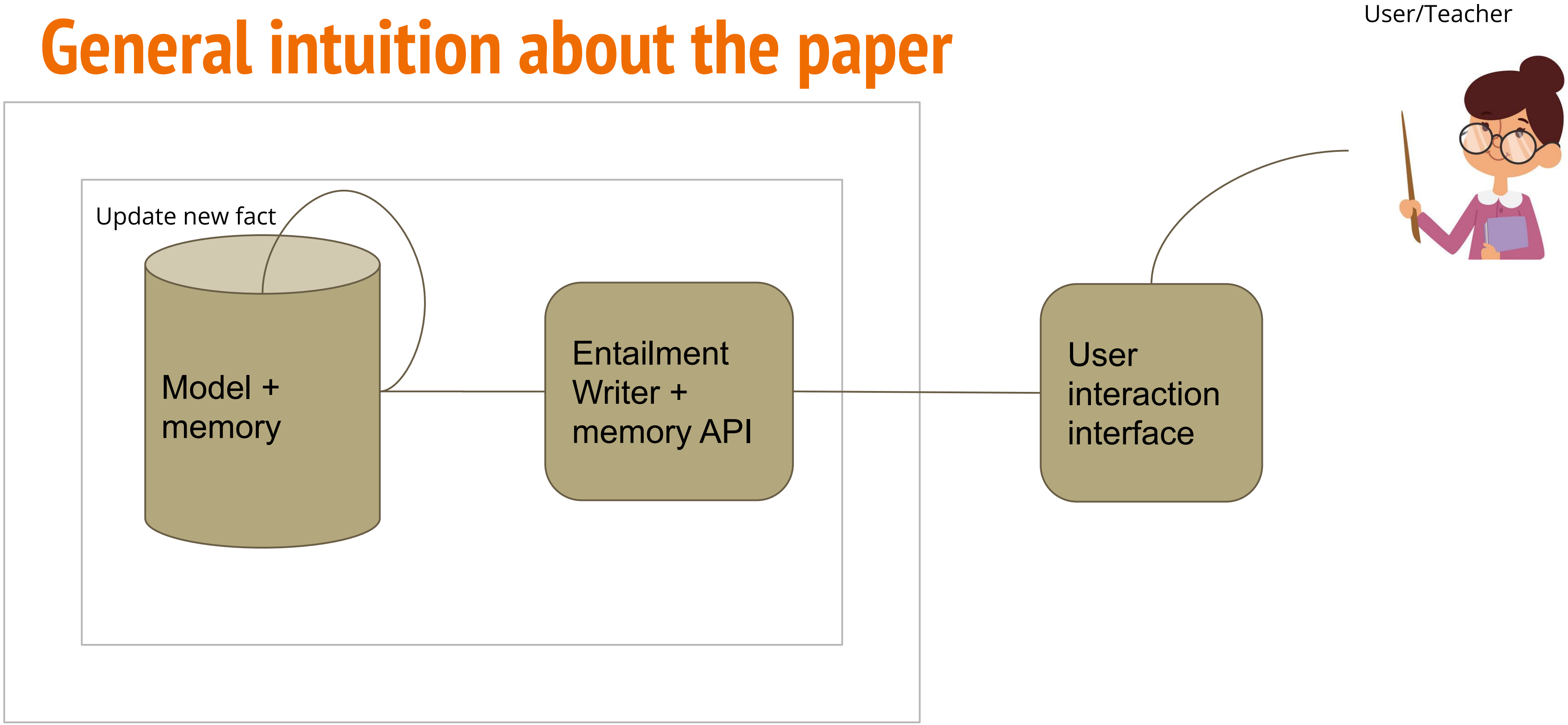
Does entailment-based reasoning lead to better answers, both in terms of explaining a model's inference and answer accuracy?

Can our dynamic memory mechanism help improve EntailmentWriter's performance on new, unseen test questions without requiring model retraining, allowing users to "teach" the system?

Intuition

It is not possible to have all of the explicit knowledge. So we can teach new facts and correct the beliefs and reasoning one-step at a time and verify line of reasoning and extract proofs by self-verification, incrementally by interacting with teacher.

General intuition about the paper



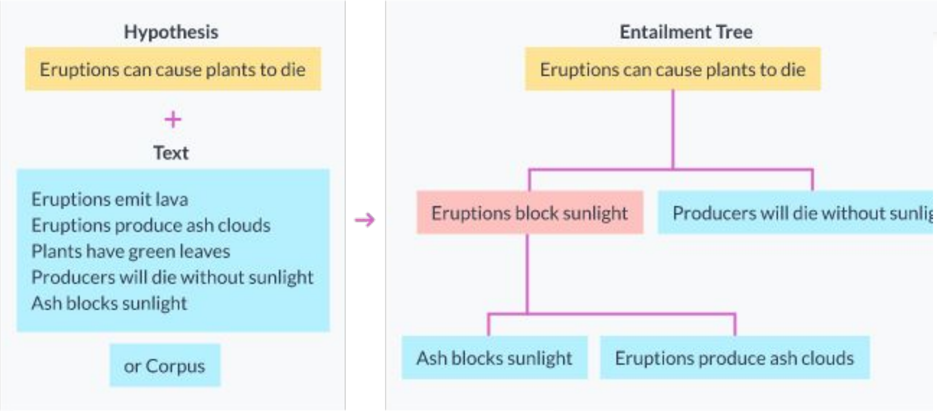
Claims

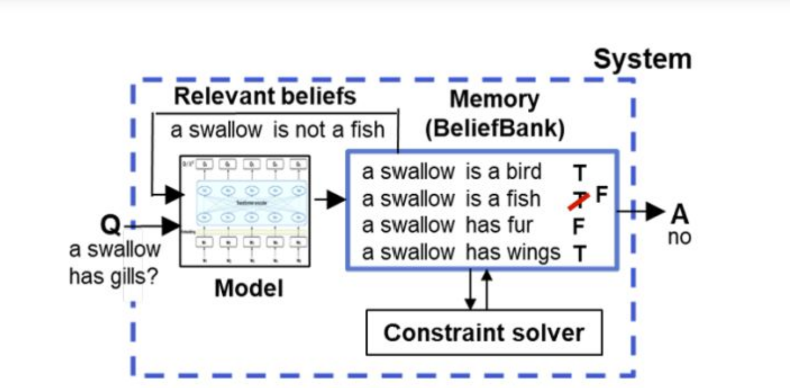
1. Faithful - answers follow from the reasoning
2. Truthful - chain reflects systems beliefs from self-querying
3. Human judgement-based evaluation: 65% clear lines of reasoning
4. Dynamic Memory

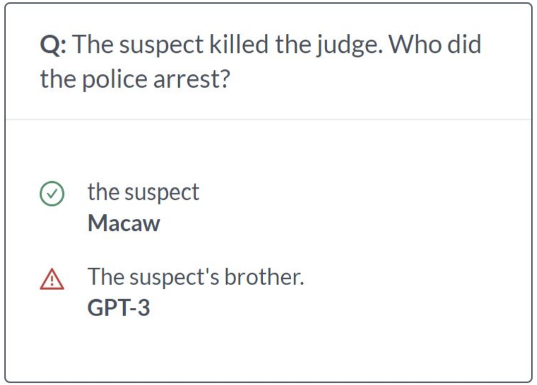
Previous works: Entailment Bank and Entailment Trees

1. Explanatory Interactive Machine Learning : Teso and Kersting, 2019
2. Explaining Answers with Entailment Trees: Dalvi, Jansen, Tafjord, Xie, Smith, Pipatanangkura and Clark, 2021
3. MACAW (“Multi-angle c(q)uestionanswering) : Tafjord and Clark, 2021
4. OBQA : Mihaylov et al., 2018
5. QuaRTz : Tafjord et al., 2019
6. Fast model editing at scale: Mitchell, Lin, Bosselut, Finn, and Manning, 2021

Previous works: Entailment Trees, Belief Bank, Macaw


Hypothesis
Eruptions can cause plants to die
+
Text
Eruptions emit lava
Eruptions produce ash clouds
Plants have green leaves
Producers will die without sunlight
Ash blocks sunlight
or Corpus
→
Entailment Tree
Eruptions can cause plants to die
Eruptions block sunlight
Producers will die without sunlig
Ash blocks sunlight
Eruptions produce ash clouds



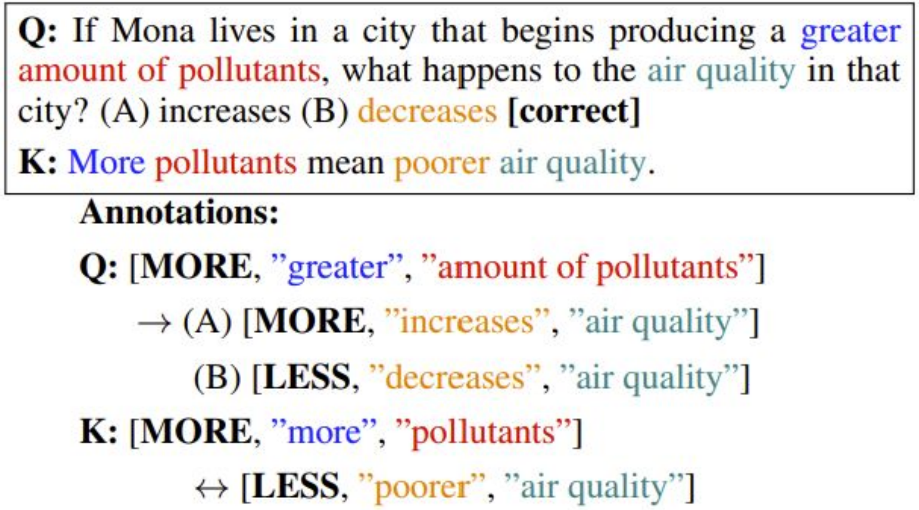


Q: The suspect killed the judge. Who did the police arrest?

✓ the suspect
Macaw

⚠ The suspect's brother.
GPT-3

Dataset QUARTZ: An Open-Domain Dataset of Qualitative Relationship



Q: If Mona lives in a city that begins producing a greater amount of pollutants, what happens to the air quality in that city? (A) increases (B) decreases **[correct]**

K: More pollutants mean poorer air quality.

Annotations:

Q: [**MORE**, ”greater”, ”amount of pollutants”]

→ (A) [**MORE**, ”increases”, ”air quality”]

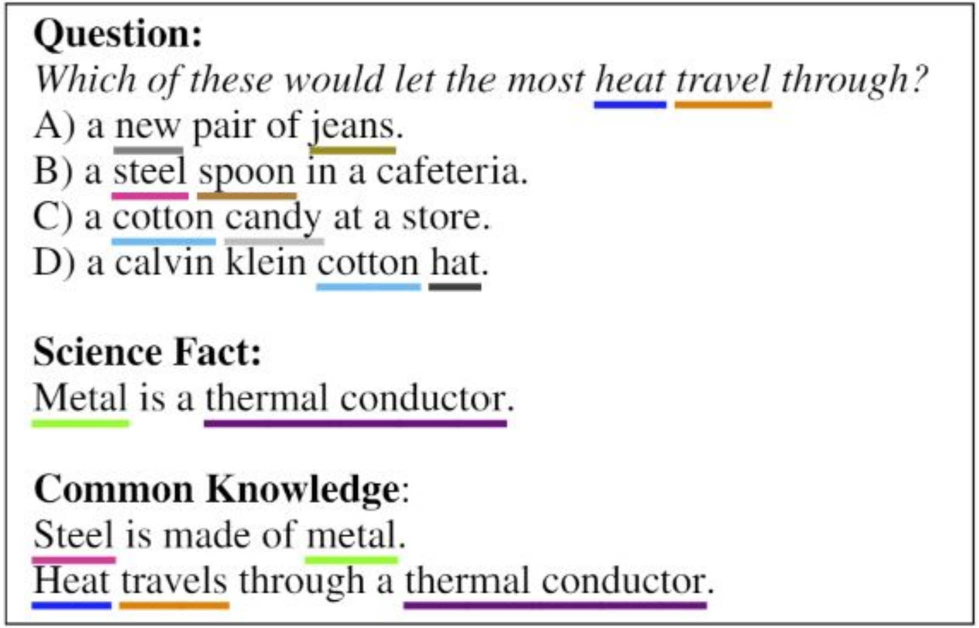
(B) [**LESS**, ”decreases”, ”air quality”]

K: [**MORE**, ”more”, ”pollutants”]

↔ [**LESS**, ”poorer”, ”air quality”]

Figure 1: QUARTZ contains situated qualitative questions, each paired with a gold background knowledge sentence and qualitative annotations.

Dataset OBQA: Open Book Question Answering



Question:

Which of these would let the most heat travel through?

A) a new pair of jeans.

B) a steel spoon in a cafeteria.

C) a cotton candy at a store.

D) a calvin klein cotton hat.

Science Fact:

Metal is a thermal conductor.

Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.

Proof Generation: Valid and Invalid cases

1. These invalid proofs will be incorrect ->
   a. fact is false or
   b. the inference is not valid entailment.
2. Slots, Values and Angles
   a. Questions
   b. Answers
   c. Multiple choices
   d. Hypothesis
   e. Premises

http://cognitiveai.org/dist/entailmentbank-book-may2022.pdf

Generating proofs

1. over-generates candidate proofs,
2. removes those that the model itself does not “believe”

Introduced approaches

Generate individual inference steps and a controller chains them together from outside loop

Backward-chaining

Explicitly state implicit knowledge

EntailmentWriter -> Edit the model

Dynamic Memory

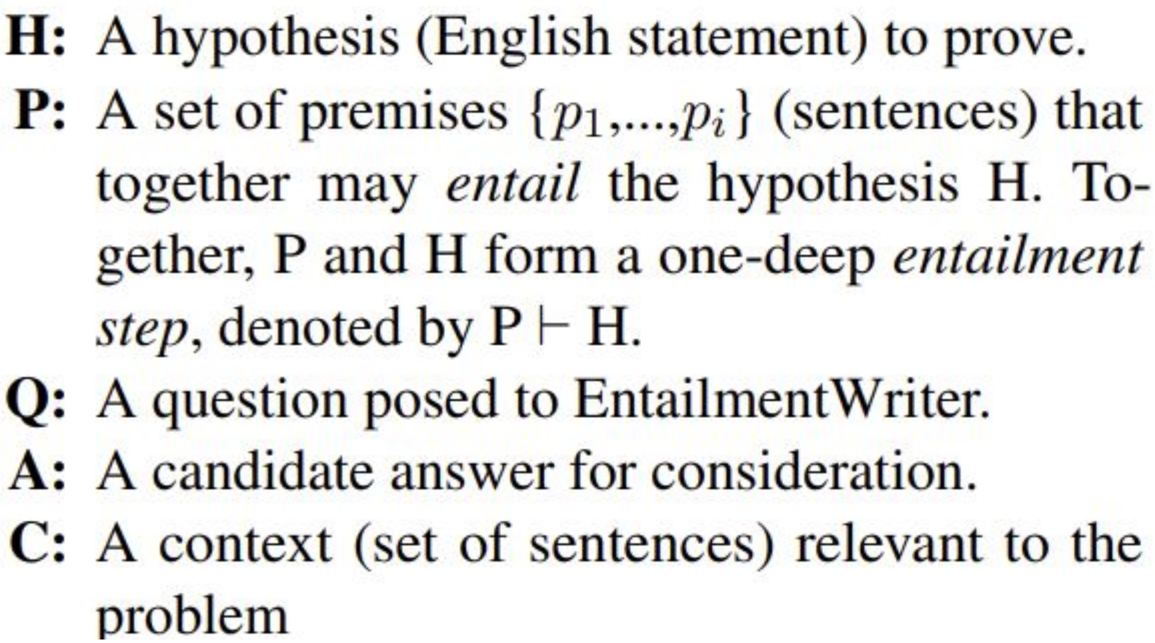
Feedback and Interaction

Memory Update - for new beliefs & facts

Memory Update (EntailmentWriter):

- (Q,{Mem,Model},A)→(Q,{Mem’,Model},A’)
- BM25 Search: How often the term occurs in all documents and penalize the most common terms
- Generate r+1 proofs but start with rth proof the topmost proof is forced

Angles - Multitask training



H: A hypothesis (English statement) to prove.

P: A set of premises $\{p_1,...,p_i\}$ (sentences) that together may *entail* the hypothesis H. Together, P and H form a one-deep *entailment step*, denoted by P $\vdash$ H.

Q: A question posed to EntailmentWriter.

A: A candidate answer for consideration.

C: A context (set of sentences) relevant to the problem

3 Angle training

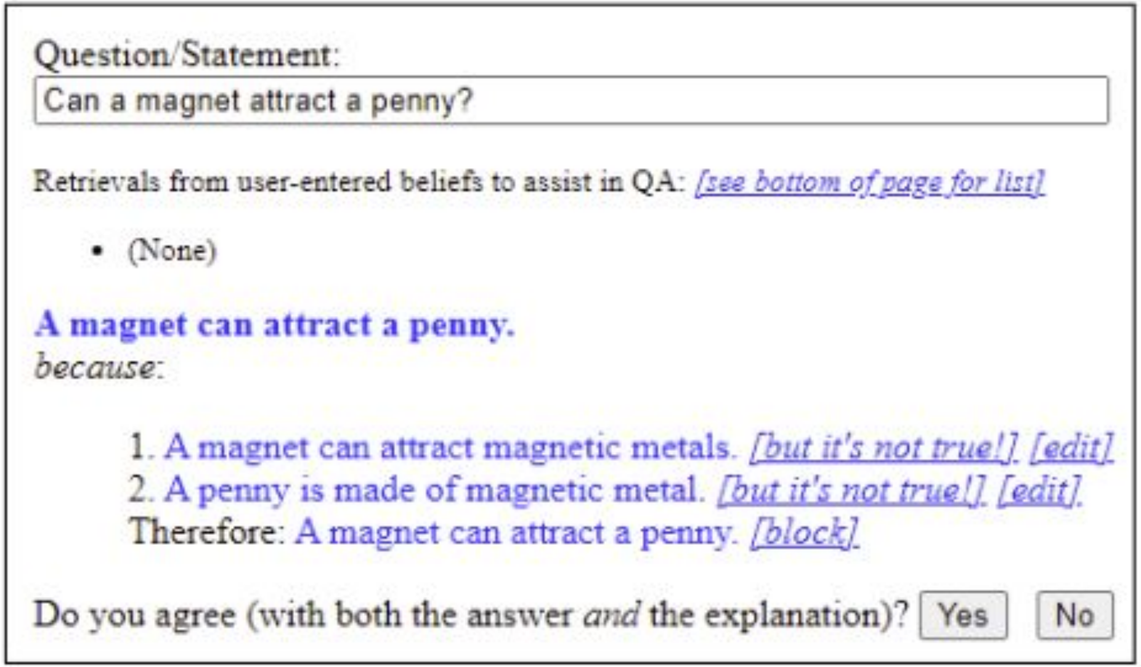
T5

H → P (proof generation),

H → Sd (fact scoring/verification), and

P H → Se (entailment scoring/verification)

Step 1 with User interaction



Question/Statement:

Can a magnet attract a penny?

Retrievals from user-entered beliefs to assist in QA: *[see bottom of page for list]*

- (None)

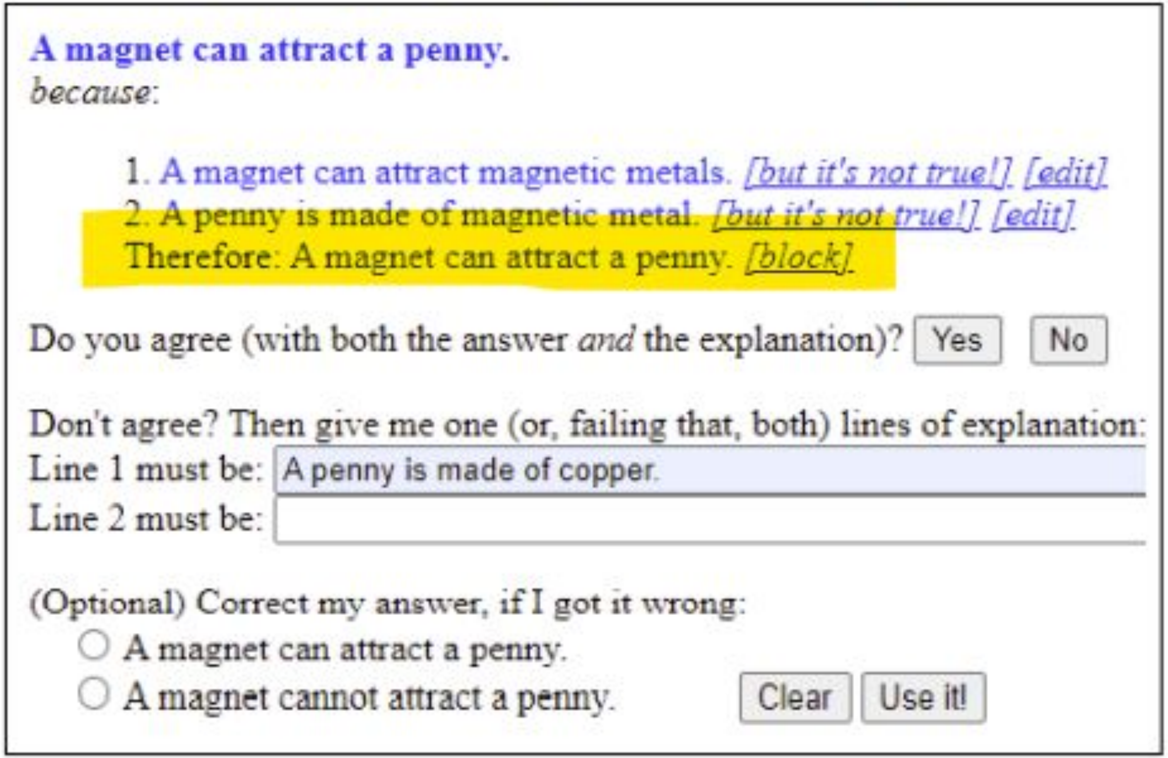
A magnet can attract a penny.
because:

1. A magnet can attract magnetic metals. *[but it's not true!]* *[edit]*
2. A penny is made of magnetic metal. *[but it's not true!]* *[edit]*

Therefore: A magnet can attract a penny. *[block]*

Do you agree (with both the answer *and* the explanation)?

Step 2 with User interaction



A magnet can attract a penny.
because:

1. A magnet can attract magnetic metals. *[but it's not true!]* *[edit]*
2. A penny is made of magnetic metal. *[but it's not true!]* *[edit]*

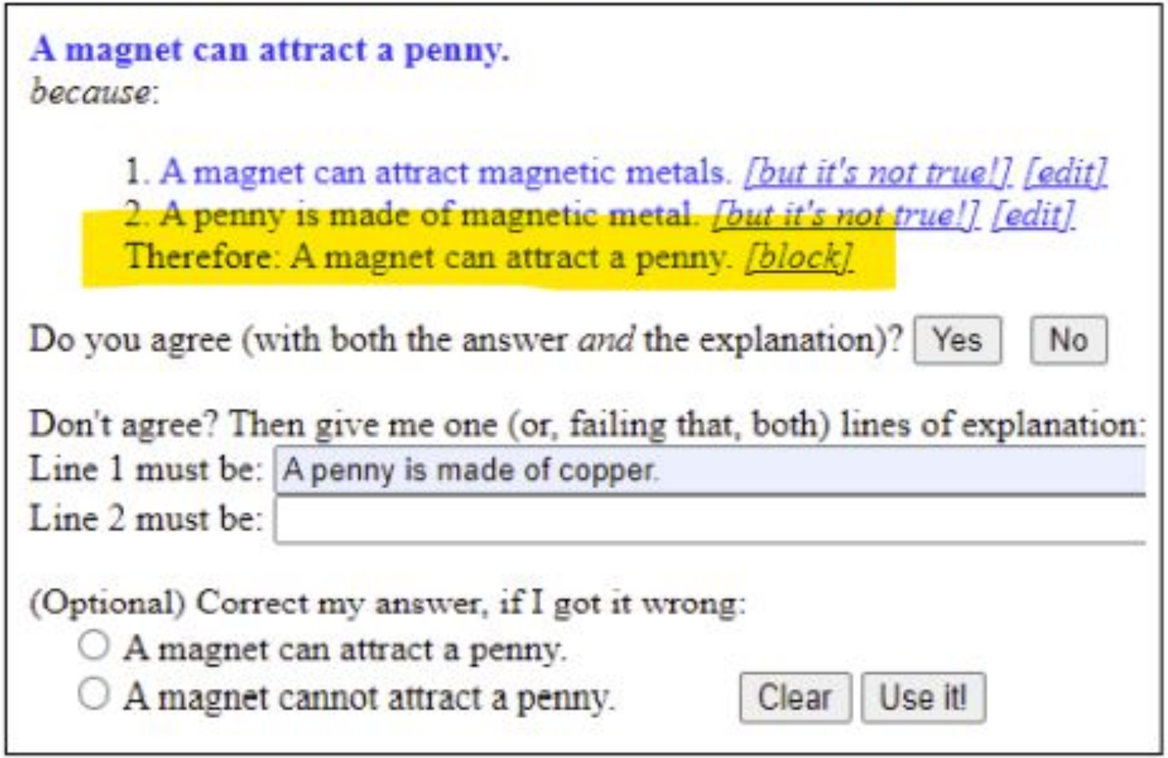
Therefore: A magnet can attract a penny. *[block]*

Do you agree (with both the answer *and* the explanation)? Yes No

Don't agree? Then give me one (or, failing that, both) lines of explanation:
Line 1 must be: A penny is made of copper.
Line 2 must be:

(Optional) Correct my answer, if I got it wrong:
○ A magnet can attract a penny.
○ A magnet cannot attract a penny.

Step 2 with User interaction



A magnet can attract a penny.
because:

1. A magnet can attract magnetic metals. *[but it's not true!]* *[edit]*
2. A penny is made of magnetic metal. *[but it's not true!]* *[edit]*

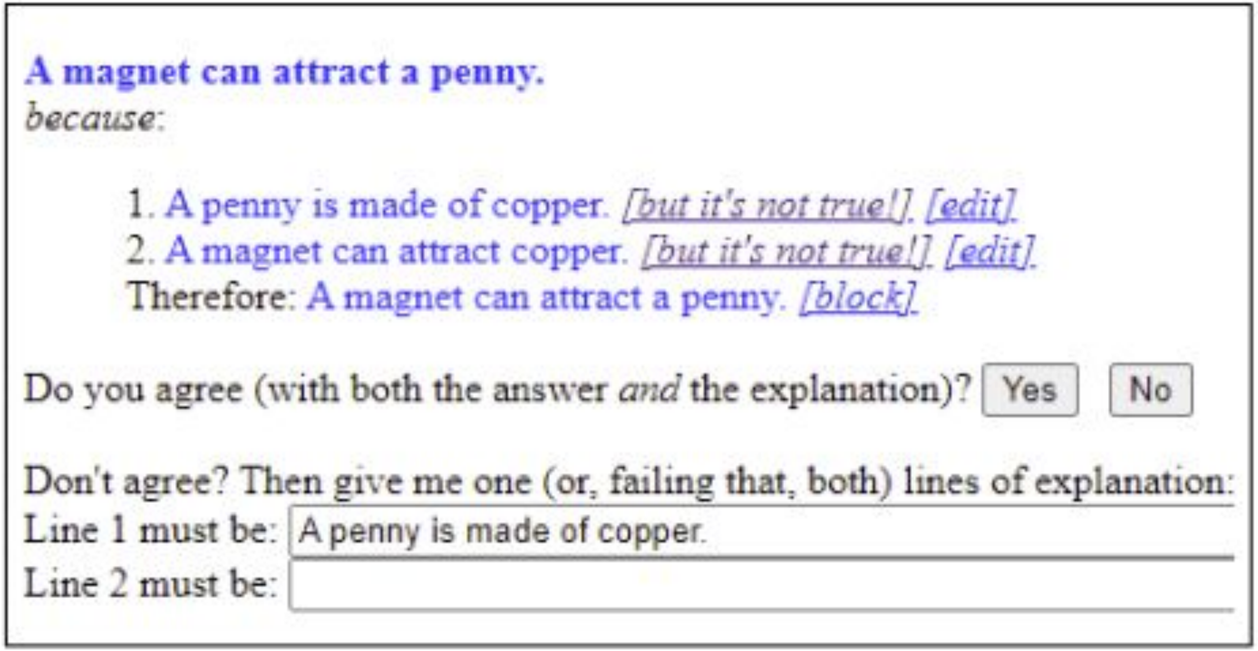
Therefore: A magnet can attract a penny. *[block]*

Do you agree (with both the answer *and* the explanation)? Yes No

Don't agree? Then give me one (or, failing that, both) lines of explanation:
Line 1 must be: A penny is made of copper.
Line 2 must be:

(Optional) Correct my answer, if I got it wrong:
○ A magnet can attract a penny.
○ A magnet cannot attract a penny.

Step 3 with User interaction



A magnet can attract a penny.
because:

1. A penny is made of copper. *[but it's not true!]* *[edit]*
2. A magnet can attract copper. *[but it's not true!]* *[edit]*

Therefore: A magnet can attract a penny. *[block]*

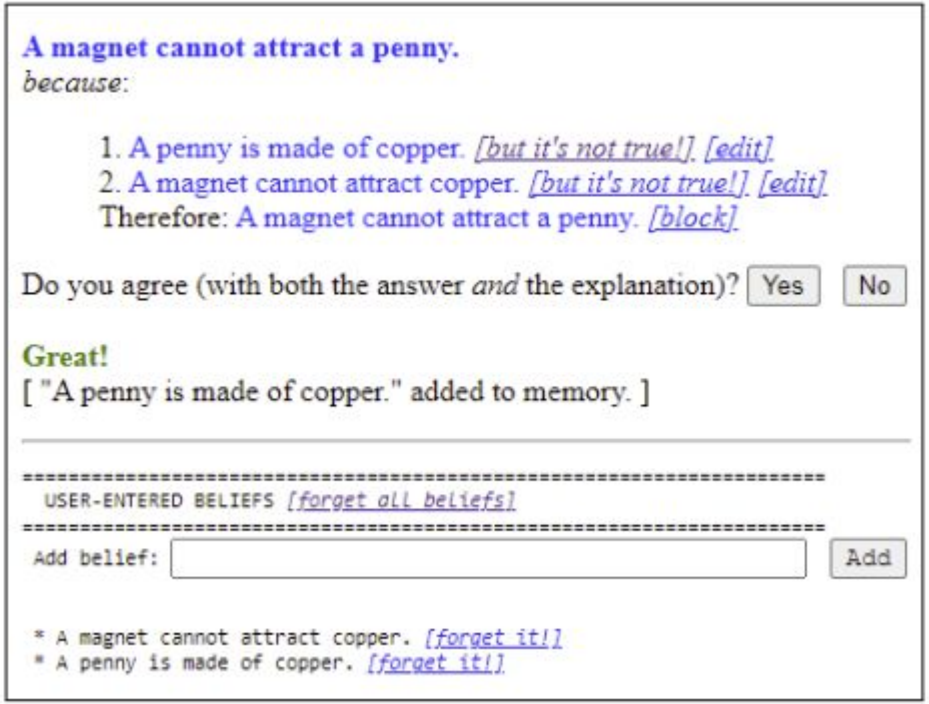
Do you agree (with both the answer *and* the explanation)?

Don't agree? Then give me one (or, failing that, both) lines of explanation:

Line 1 must be: A penny is made of copper.

Line 2 must be:

Step 4 with User interaction



A magnet cannot attract a penny.
because:

1. A penny is made of copper. *[but it's not true!]* *[edit]*
2. A magnet cannot attract copper. *[but it's not true!]* *[edit]*

Therefore: A magnet cannot attract a penny. *[block]*

Do you agree (with both the answer *and* the explanation)?

Great!
["A penny is made of copper." added to memory.]

```
======================================================================
  USER-ENTERED BELIEFS [forget all beliefs]
======================================================================
 Add belief:


 * A magnet cannot attract copper. [forget it!]
 * A penny is made of copper. [forget it!]
```

Users' Assessment

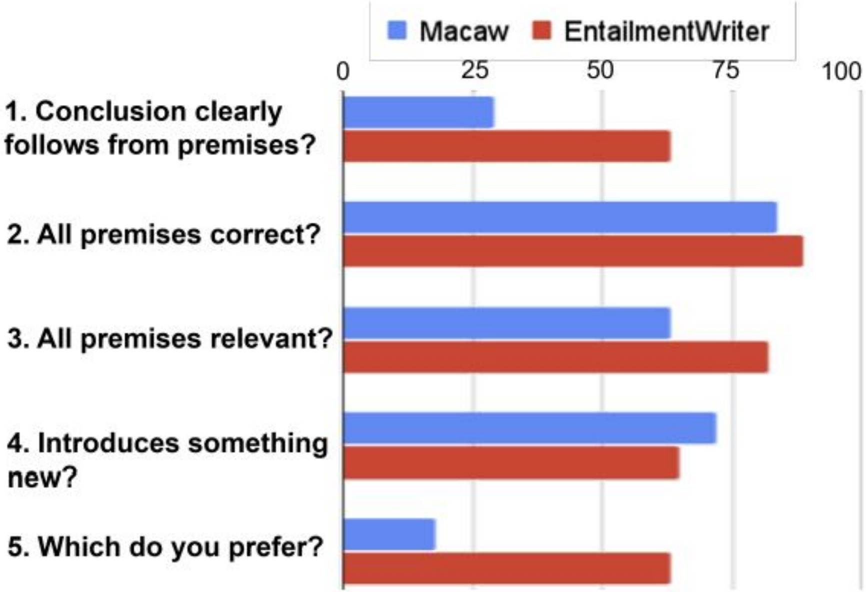


Figure 6: Users' assessments of EntailmentWriter's proofs (red), compared to Macaw's explanations (blue), showing percent of times annotators answered "Yes". EntailmentWriter's conclusions were judged to "clearly follow from the premises" in over 65% of the proofs (first bar), substantially more than Macaw's explanations (30%).

Results over OBQA and QUaRTz

Direct QA - zero shot over Entailment Writer
And uses only DirectQA fact scoring/verification angle

Entailment Writer UpperBound: EntailmentWriter's memory is set to all the core facts used in all training questions

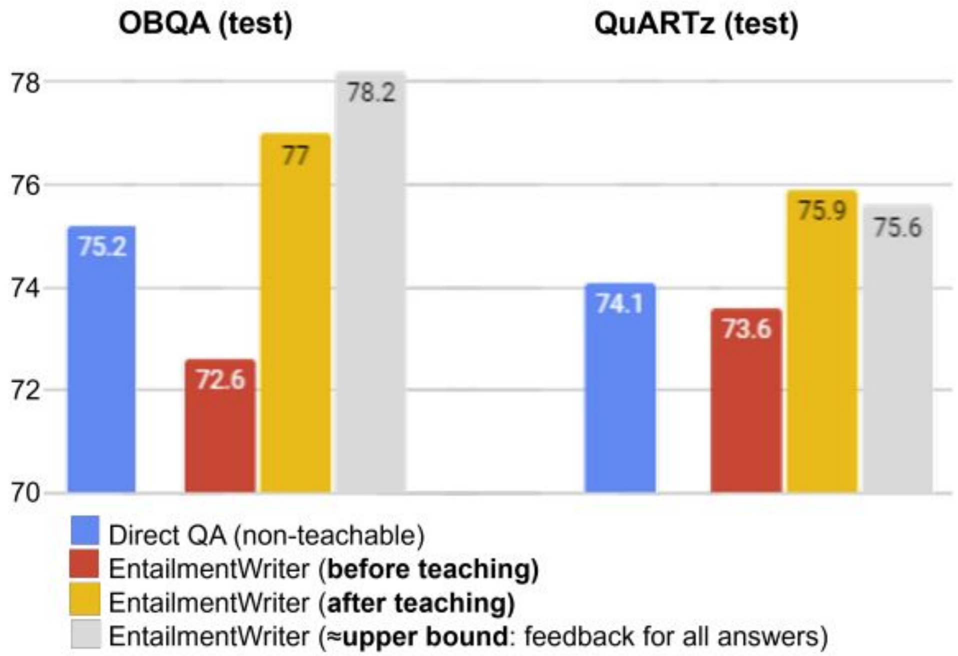


Figure 7: EntailmentWriter's performance on the hidden test sets improves with user feedback, improving over direct QA and coming close (within ≈ 1%) of the upper bound of using feedback on all answers.

Baseline vs EntailmentWriter's performance

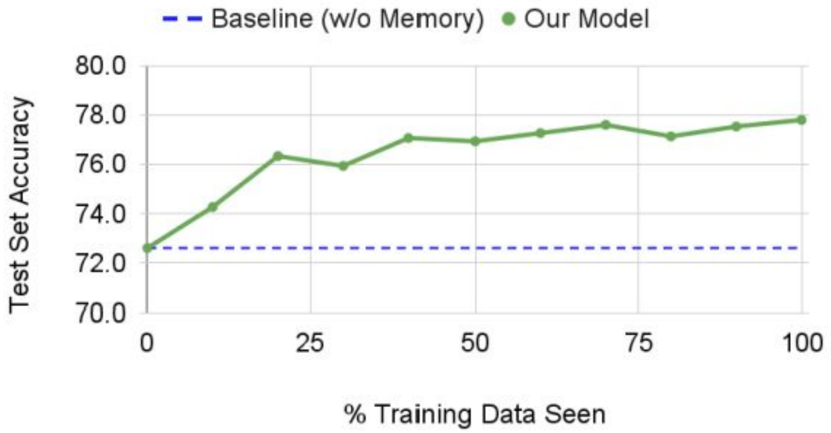
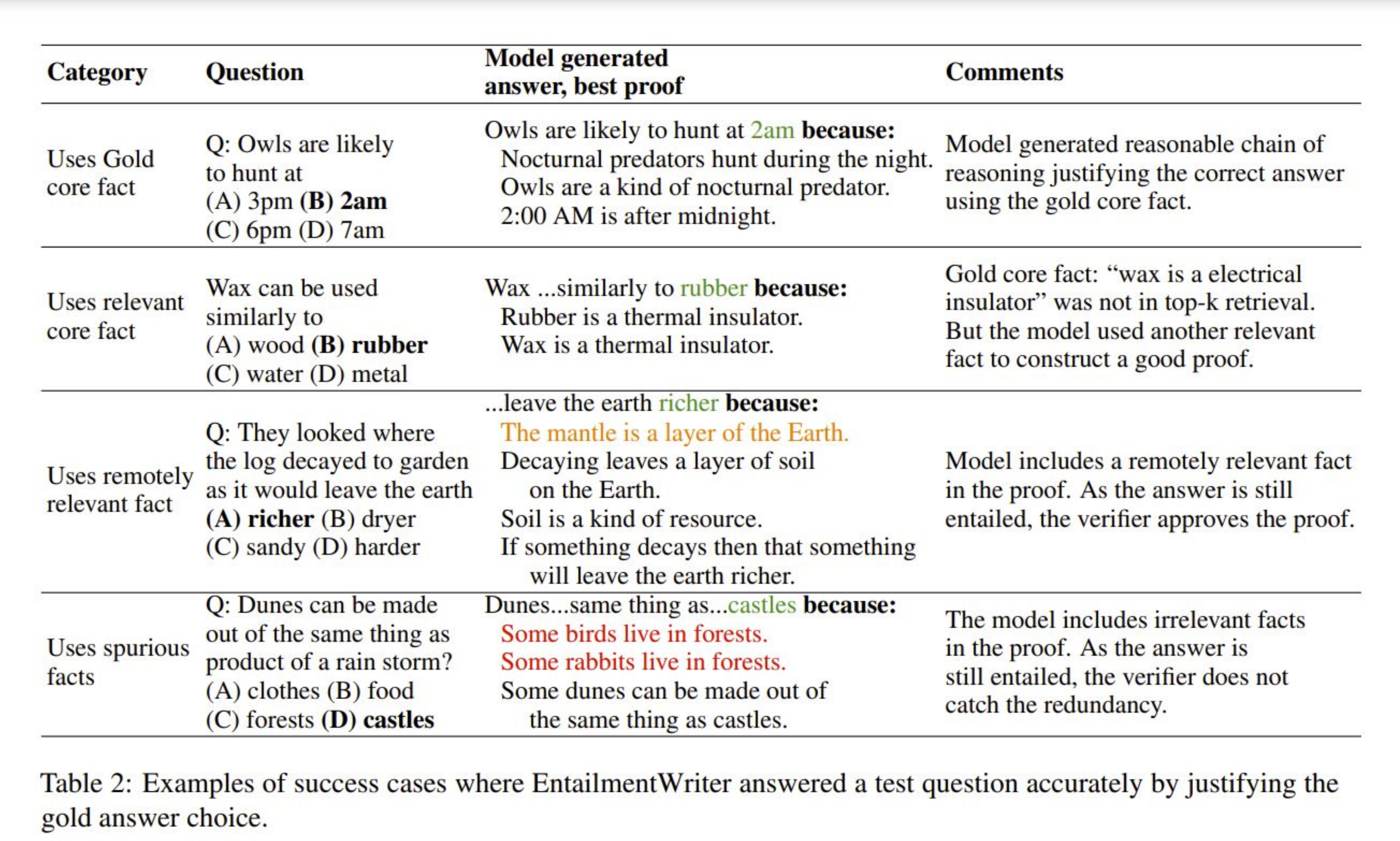


Figure 8: EntailmentWriter's performance on OBQA test improves as it sees a larger fraction of training data and stores feedback for wrong answers in its memory.

Success Examples



Category	Question	Model generated answer, best proof	Comments
Uses Gold core fact	Q: Owls are likely to hunt at (A) 3pm **(B) 2am** (C) 6pm (D) 7am	Owls are likely to hunt at 2am **because:** Nocturnal predators hunt during the night. Owls are a kind of nocturnal predator. 2:00 AM is after midnight.	Model generated reasonable chain of reasoning justifying the correct answer using the gold core fact.
Uses relevant core fact	Wax can be used similarly to (A) wood **(B) rubber** (C) water (D) metal	Wax ...similarly to rubber **because:** Rubber is a thermal insulator. Wax is a thermal insulator.	Gold core fact: "wax is a electrical insulator" was not in top-k retrieval. But the model used another relevant fact to construct a good proof.
Uses remotely relevant fact	Q: They looked where the log decayed to garden as it would leave the earth **(A) richer** (B) dryer (C) sandy (D) harder	...leave the earth richer **because:** The mantle is a layer of the Earth. Decaying leaves a layer of soil on the Earth. Soil is a kind of resource. If something decays then that something will leave the earth richer.	Model includes a remotely relevant fact in the proof. As the answer is still entailed, the verifier approves the proof.
Uses spurious facts	Q: Dunes can be made out of the same thing as product of a rain storm? (A) clothes (B) food (C) forests **(D) castles**	Dunes...same thing as...castles **because:** Some birds live in forests. Some rabbits live in forests. Some dunes can be made out of the same thing as castles.	The model includes irrelevant facts in the proof. As the answer is still entailed, the verifier does not catch the redundancy.

Table 2: Examples of success cases where EntailmentWriter answered a test question accurately by justifying the gold answer choice.

Assessment of Correctness of EntailmentWriter

entailment verifier - has fewer examples.

Spurious fact changes previously correctly written fact

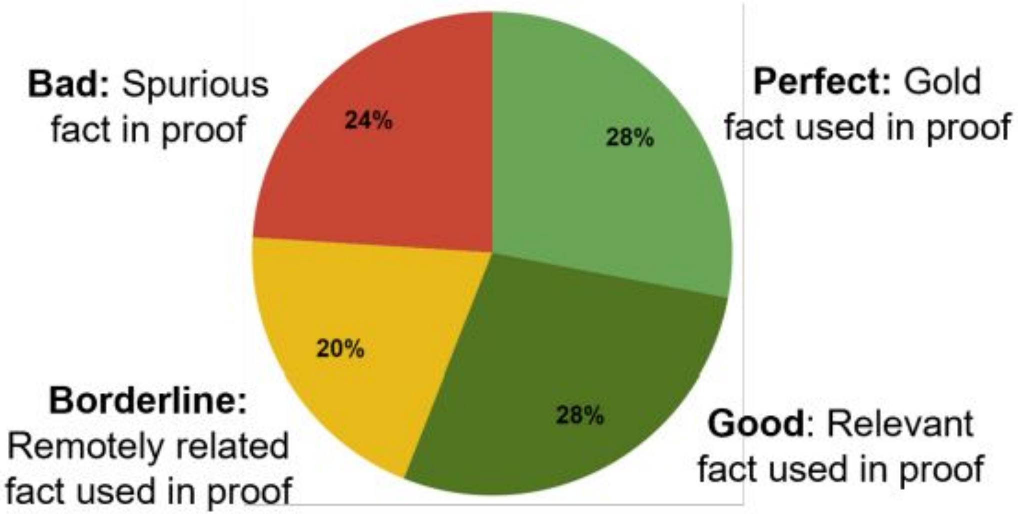
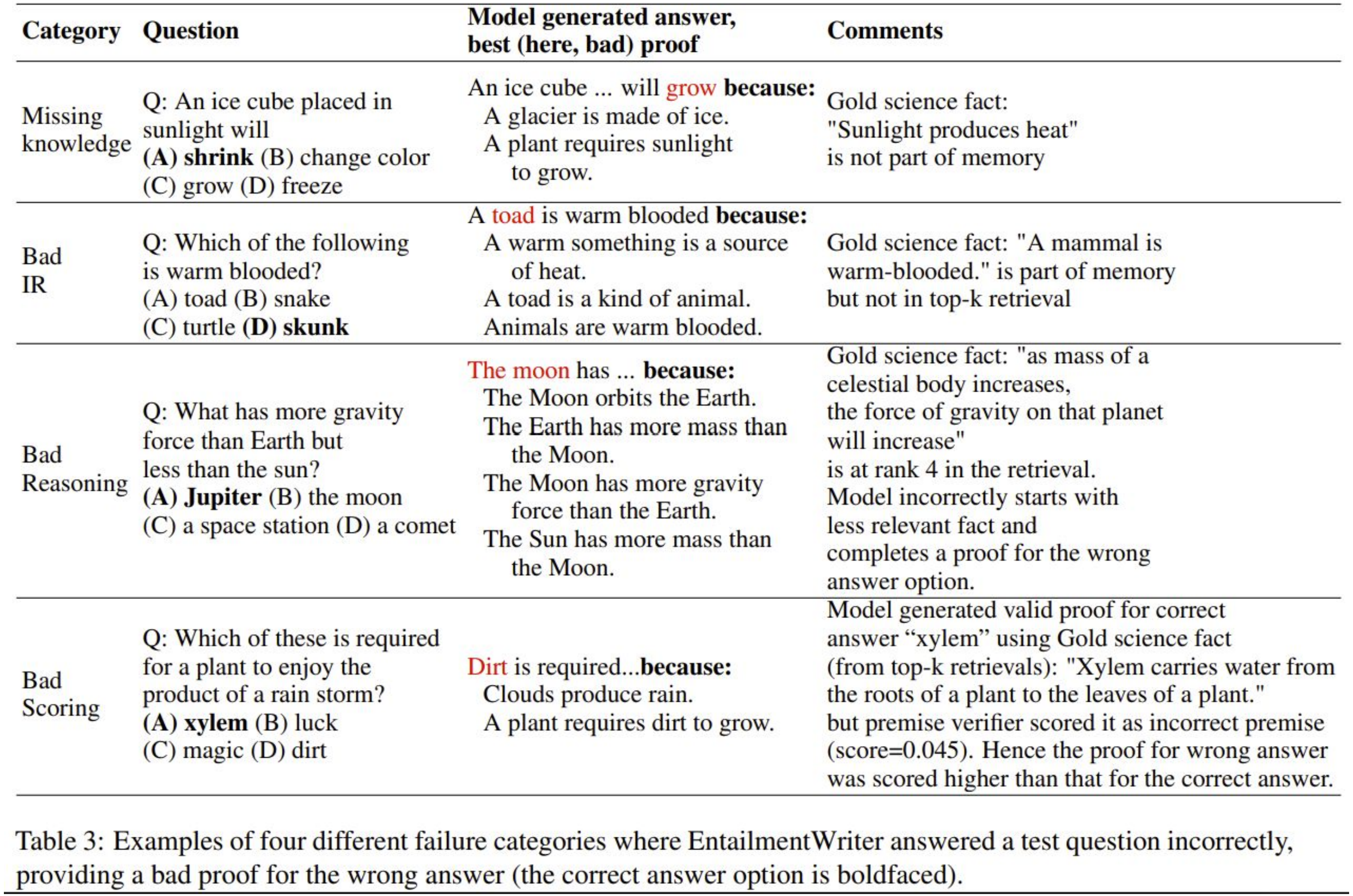


Figure 9: EntailmentWriter was right for the right reasons in ≈75% of its correct answers. (Examples are shown in Table 2).

**Github code is not available to verify this. **

Failure analysis



Category	Question	Model generated answer, best (here, bad) proof	Comments
Missing knowledge	Q: An ice cube placed in sunlight will **(A) shrink** (B) change color (C) grow (D) freeze	An ice cube ... will grow **because:** A glacier is made of ice. A plant requires sunlight to grow.	Gold science fact: "Sunlight produces heat" is not part of memory
Bad IR	Q: Which of the following is warm blooded? (A) toad (B) snake (C) turtle **(D) skunk**	A toad is warm blooded **because:** A warm something is a source of heat. A toad is a kind of animal. Animals are warm blooded.	Gold science fact: "A mammal is warm-blooded." is part of memory but not in top-k retrieval
Bad Reasoning	Q: What has more gravity force than Earth but less than the sun? **(A) Jupiter** (B) the moon (C) a space station (D) a comet	The moon has ... **because:** The Moon orbits the Earth. The Earth has more mass than the Moon. The Moon has more gravity force than the Earth. The Sun has more mass than the Moon.	Gold science fact: "as mass of a celestial body increases, the force of gravity on that planet will increase" is at rank 4 in the retrieval. Model incorrectly starts with less relevant fact and completes a proof for the wrong answer option.
Bad Scoring	Q: Which of these is required for a plant to enjoy the product of a rain storm? **(A) xylem** (B) luck (C) magic (D) dirt	Dirt is required...**because:** Clouds produce rain. A plant requires dirt to grow.	Model generated valid proof for correct answer “xylem” using Gold science fact (from top-k retrievals): "Xylem carries water from the roots of a plant to the leaves of a plant." but premise verifier scored it as incorrect premise (score=0.045). Hence the proof for wrong answer was scored higher than that for the correct answer.

Table 3: Examples of four different failure categories where EntailmentWriter answered a test question incorrectly, providing a bad proof for the wrong answer (the correct answer option is boldfaced).

Assessment Of Failure

EntailmentWriter did not choose correct proof even if it exists.

Score was higher for a different proof than expected proof.

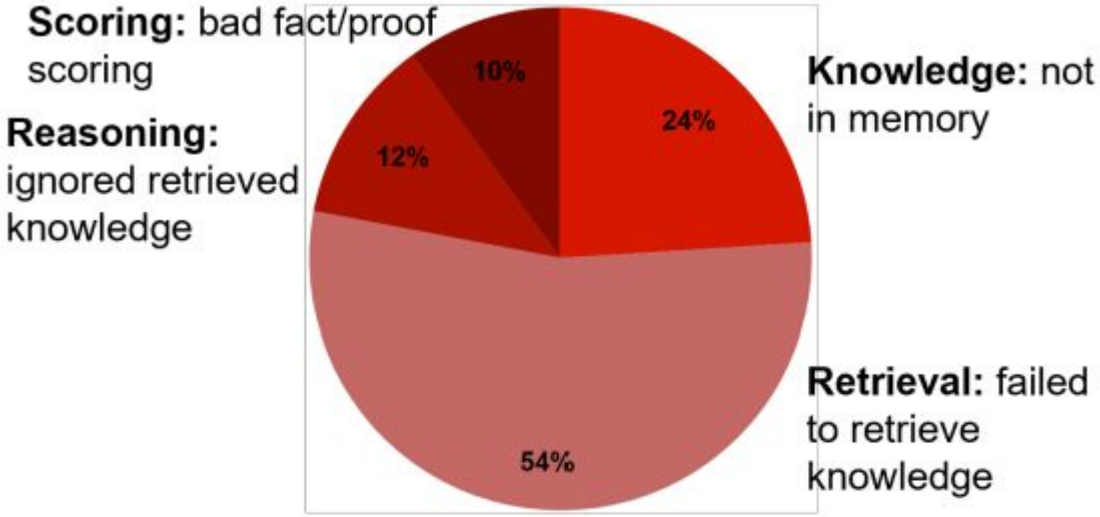


Figure 10: Causes of failure (%) for EntailmentWriter's incorrect answers. (Examples are shown in Table 3).

Hidden test set

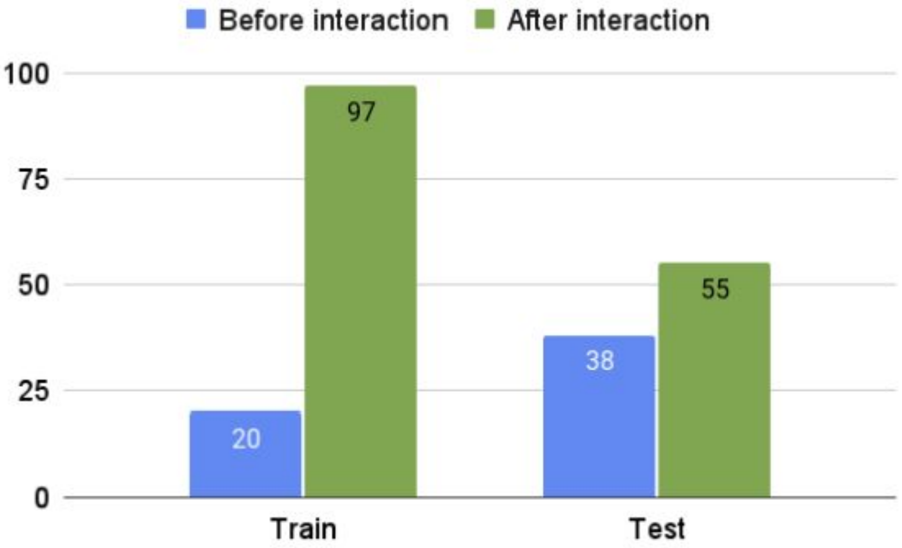


Figure 11: EntailmentWriter's performance (% correct) substantially improves on a hidden test set (from 38% to 55%), a subset of OBQA, after users correct/expand its knowledge for the training questions. (Results are averaged over 8 users).

Interesting concepts : Simulated Teacher

Simulated Teacher to validate proofs

- gold entailment proof as user feedback
- Model is frozen in this simulated user scenario
- Only growing the Dynamic memory with new contents

Interesting concept: Cognitive Architecture

Makes use of Memory

Oversimplified and called as “Simple Cognitive Architecture”

Seems to be inspired by Connectionist models with Memory

Student-Teacher learning

**** Few are already cited in paper:** Active & co-active learning (interactiveness)

More tests need to be conducted

Given that Explicit knowledge was not used:

How does the model perform over challenge datasets in-memory?

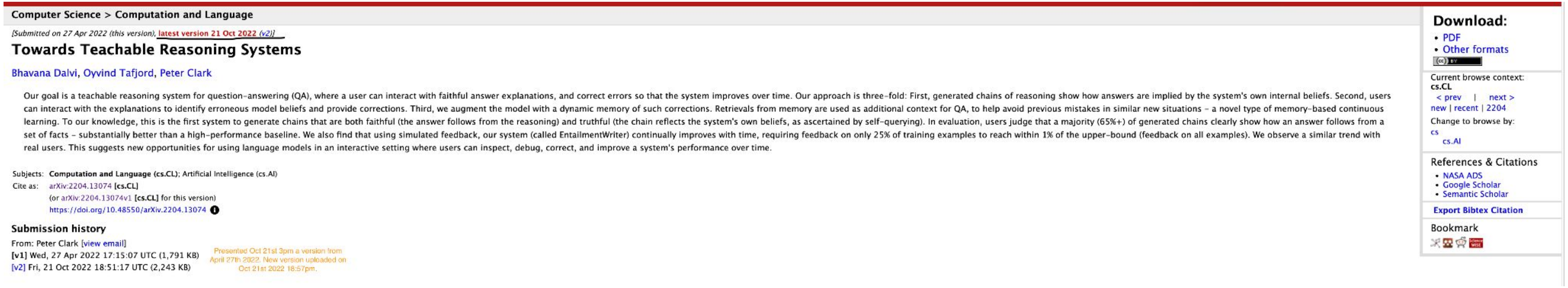
- Adversarial
- Negation
- irrelevant facts

How does the model perform over challenge facts **not** in the memory?

How does model respond to unavailable proofs?

Can we introduce another Teacher model to teach and provide feedback?

New version released by AllenAI on Oct 21st evening



Computer Science > Computation and Language

[Submitted on 27 Apr 2022 (this version), latest version 21 Oct 2022 (v2)]

Towards Teachable Reasoning Systems

Bhavana Dalvi, Oyvind Tafjord, Peter Clark

Our goal is a teachable reasoning system for question-answering (QA), where a user can interact with faithful answer explanations, and correct errors so that the system improves over time. Our approach is three-fold: First, generated chains of reasoning show how answers are implied by the system's own internal beliefs. Second, users can interact with the explanations to identify erroneous model beliefs and provide corrections. Third, we augment the model with a dynamic memory of such corrections. Retrievals from memory are used as additional context for QA, to help avoid previous mistakes in similar new situations - a novel type of memory-based continuous learning. To our knowledge, this is the first system to generate chains that are both faithful (the answer follows from the reasoning) and truthful (the chain reflects the system's own beliefs, as ascertained by self-querying). In evaluation, users judge that a majority (65%+) of generated chains clearly show how an answer follows from a set of facts - substantially better than a high-performance baseline. We also find that using simulated feedback, our system (called EntailmentWriter) continually improves with time, requiring feedback on only 25% of training examples to reach within 1% of the upper-bound (feedback on all examples). We observe a similar trend with real users. This suggests new opportunities for using language models in an interactive setting where users can inspect, debug, correct, and improve a system's performance over time.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)
Cite as: arXiv:2204.13074 **[cs.CL]**
(or arXiv:2204.13074v1 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2204.13074

Submission history

From: Peter Clark [view email]
[v1] Wed, 27 Apr 2022 17:15:07 UTC (1,791 KB)
[v2] Fri, 21 Oct 2022 18:51:17 UTC (2,243 KB)

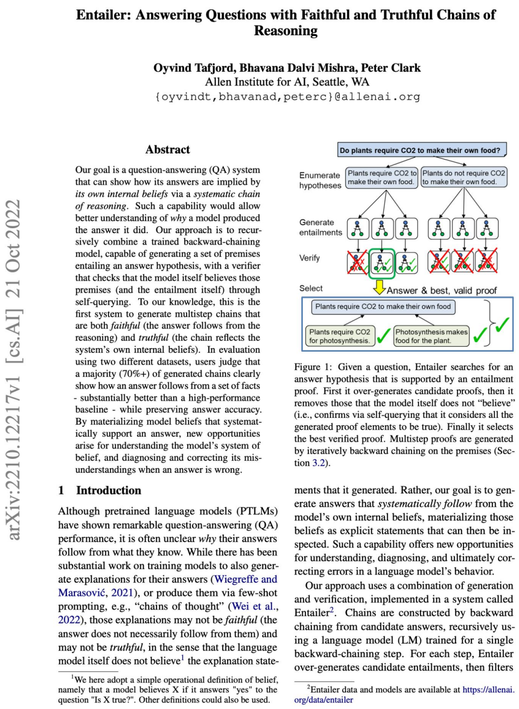
Presented Oct 21st 3pm a version from April 27th 2022. New version uploaded on Oct 21st 2022 18:57pm.

Added a Teachable reasoning architecture (which was not present before the presentation).

New Paper : Entailer : Oct 26th 2022 follow up

Similar to Towards Teachable reasoning systems

Does not use “Context” C



Entailer: Answering Questions with Faithful and Truthful Chains of Reasoning

Oyvind Tafjord, Bhavana Dalvi Mishra, Peter Clark
Allen Institute for AI, Seattle, WA
{oyvindt,bhavanad,peterc}@allenai.org

Abstract

Our goal is a question-answering (QA) system that can show how its answers are implied by *its own internal beliefs* via a *systematic chain of reasoning*. Such a capability would allow better understanding of *why* a model produced the answer it did. Our approach is to recursively combine a trained backward-chaining model, capable of generating a set of premises entailing an answer hypothesis, with a verifier that checks that the model itself believes those premises (and the entailment itself) through self-querying. To our knowledge, this is the first system to generate multistep chains that are both *faithful* (the answer follows from the reasoning) and *truthful* (the chain reflects the system's own internal beliefs). In evaluation using two different datasets, users judge that a majority (70%+) of generated chains clearly show how an answer follows from a set of facts - substantially better than a high-performance baseline - while preserving answer accuracy. By materializing model beliefs that systematically support an answer, new opportunities arise for understanding the model's system of belief, and diagnosing and correcting its misunderstandings when an answer is wrong.

Do plants require CO2 to make their own food?
Enumerate hypotheses
Plants require CO2 to make their own food.
Plants do not require CO2 to make their own food.
Generate entailments
Verify
Select
Answer & best, valid proof
Plants require CO2 to make their own food
Plants require CO2 for photosynthesis.
Photosynthesis makes food for the plant.

Figure 1: Given a question, Entailer searches for an answer hypothesis that is supported by an entailment proof. First it over-generates candidate proofs, then it removes those that the model itself does not “believe” (i.e., confirms via self-querying that it considers all the generated proof elements to be true). Finally it selects the best verified proof. Multistep proofs are generated by iteratively backward chaining on the premises (Section 3.2).

1 Introduction

Although pretrained language models (PTLMs) have shown remarkable question-answering (QA) performance, it is often unclear *why* their answers follow from what they know. While there has been substantial work on training models to also generate explanations for their answers (Wiegreffe and Marasović, 2021), or produce them via few-shot prompting, e.g., “chains of thought” (Wei et al., 2022), those explanations may not be *faithful* (the answer does not necessarily follow from them) and may not be *truthful*, in the sense that the language model itself does not believe[1] the explanation statements that it generated. Rather, our goal is to generate answers that *systematically follow* from the model's own internal beliefs, materializing those beliefs as explicit statements that can then be inspected. Such a capability offers new opportunities for understanding, diagnosing, and ultimately correcting errors in a language model's behavior.

Our approach uses a combination of generation and verification, implemented in a system called Entailer[2]. Chains are constructed by backward chaining from candidate answers, recursively using a language model (LM) trained for a single backward-chaining step. For each step, Entailer over-generates candidate entailments, then filters

[1] We here adopt a simple operational definition of belief, namely that a model believes X if it answers "yes" to the question "Is X true?". Other definitions could also be used.

[2] Entailer data and models are available at https://allenai.org/data/entailer