Overview

- Large language models have been shown to struggle with limited context memory and multi-step reasoning
- We propose a method for solving both of these problems by allowing the model to ask and answer questions
- The model can deviate from the input context at any time for self-questioning
- This allows the model to recall information and reason on the fly as it reads context, extending its memory and enabling multi-step reasoning

Multi-hop reasoning

Given "Alice has the box."

"Alice is at the park." one can infer "The box is at the park."

which can be further combined with a later statement "The key is in the box." to conclude that "The key is at the park.".

State-tracking

Assume x=5 initially, and then x gets incremented by 1. If the model correctly writes x=6, it can safely remove the original x=5 statement from its context.

Method: Self-questioning

2 hop question:
Bob has the key. Bob is at the park.
Q: where is the key? The key is at the park

3 hop question:
The ball is in the bag. Alice has the bag. Alice is at the lake.
Q: where is the ball? The ball is at the lake

3 hop question with self-questioning:
The ball is in the bag. Alice has the bag. SQ: who has the ball? Alice has the ball. Alice is at the lake.
Q: where is the ball? The ball is at the lake

Data

Algorithmic Task

Original context:
```
e = 3; e += 1; e+=; if (e > -e) { e = 3; } g = 3;
```

Context with self-QA:
```
e = 3; print e = 3; print e = 4; print i = 4; g = 3;
```

Question: print e
Answer: e = 2;

Toy-story Task

Original context:
The banana is inside the box. Jessie has the bag.
The ball is inside the box.
The key is inside the suitcase.
```
Sid has the box.
Buzz has the suitcase.
Woody is at the station.
```

Context with self-QA:
The banana is inside the box. Jessie has the bag.
The ball is inside the box.
The key is inside the suitcase.
```
Sid has the box.
Buzz has the suitcase.
Woody is at the station.
```

Question: who has the key?
Answer: Buzz has the key.

Example output

Green tokens are the self-QAs. Red tokens are the final answer. "print" and "SQ." are special tokens to generate a self-QA for the Algorithmic and Toy-story, respectively. The model predicts the special token as the next token, it is allowed to generate a question and answer before returning to the original context.

---

tl;dr - a general method that allows language models to take internal notes in the form of self-QAs