ALiiCE: Evaluating Positional Fine-grained Citation Generation

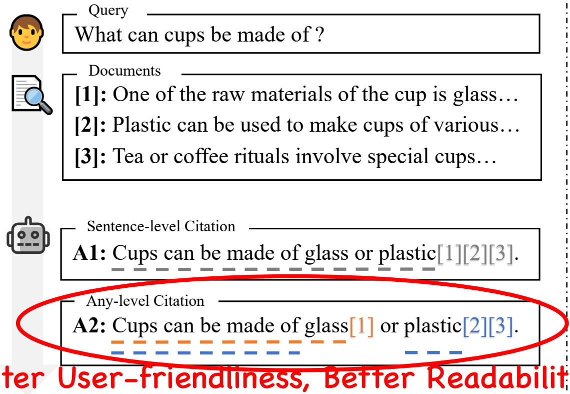
Yilong Xu[1,2,3] Jinhua Gao[1,2*] Xiaoming Yu[1,2] Baolong Bi[1,2,3] Huawei Shen[1,2,3] Xueqi Cheng[1,2,3]

[1]State Key Lab of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences

[2]Key Lab of AI Safety, Chinese Academy of Sciences [3]University of Chinese Academy of Sciences

Why need fine-grained citations?

- A sentence might not be the smallest unit capable of representing an atomic claim, potentially leading to inaccurate evaluations.
- The generated text scope of a single in-line citation often brings ambiguity, which is more common in sentence-level citations.



Query

What can cups be made of ?

Documents

[1]: One of the raw materials of the cup is glass…

[2]: Plastic can be used to make cups of various…

[3]: Tea or coffee rituals involve special cups…

Sentence-level Citation

A1: Cups can be made of glass or plastic[1][2][3].

Any-level Citation

A2: Cups can be made of glass[1] or plastic[2][3].

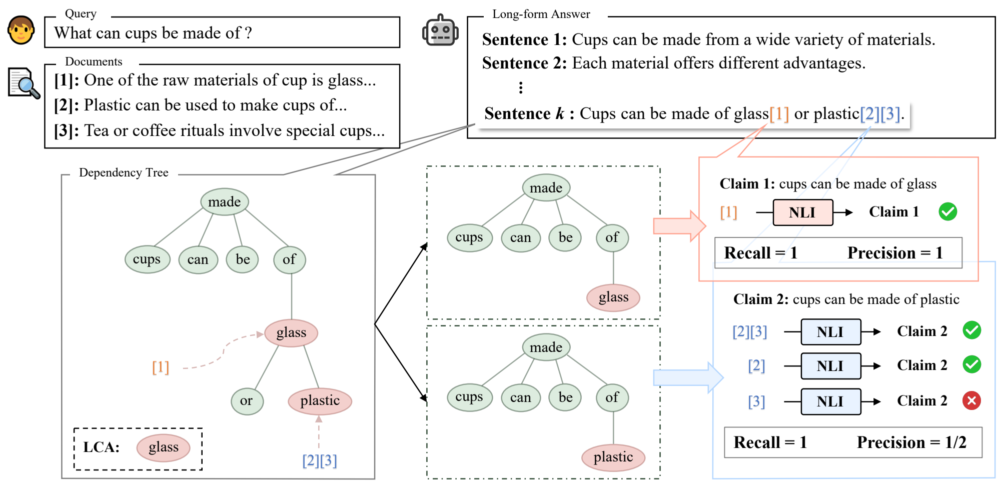
Better User-friendliness, Better Readability!!

Figure 1: “Sentence-level” vs. “Any-level” in the task of citation text generation.

We propose this improved task, called Positional Fine-grained Citation Text Generation, but there is no effective method to evaluate it.

So we introduce ALiiCE to fill this gap.

ALiiCE’s pipeline



Query

What can cups be made of ?

Documents

[1]: One of the raw materials of cup is glass...

[2]: Plastic can be used to make cups of...

[3]: Tea or coffee rituals involve special cups...

Long-form Answer

Sentence 1: Cups can be made from a wide variety of materials.

Sentence 2: Each material offers different advantages.

⋮

Sentence *k* : Cups can be made of glass[1] or plastic[2][3].

Dependency Tree

Claim 1: cups can be made of glass

[1] → NLI → Claim 1 ✓

Recall = 1 **Precision = 1**

Claim 2: cups can be made of plastic

[2][3] → NLI → Claim 2 ✓

[2] → NLI → Claim 2 ✓

[3] → NLI → Claim 2 ✗

Recall = 1 **Precision = 1/2**

Figure 2: An example of how ALiiCE evaluates fine-grained citations.

- ALiCE employs a Dependency Tree based approach to parse atomic claims of each citation in the response.

ALiiCE’s metrics on citation quality

- Positional Fine-grained Citation Recall
- Positional Fine-grained Citation Precision
- Coefficient of Variation of Citation Positions
  - CVCP measures the dispersion of citation positions within a sentence.

$$\sigma(s_k) = \sqrt{\frac{1}{t}\sum_{j=1}^{t}(p_j-\mu_k)^2} \qquad CV_{CP}(\mathcal{R}) = \frac{1}{n}\sum_{k=1}^{n}\frac{\sigma(s_k)}{\mu_k}$$

Comparison results

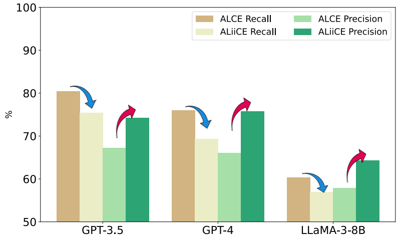
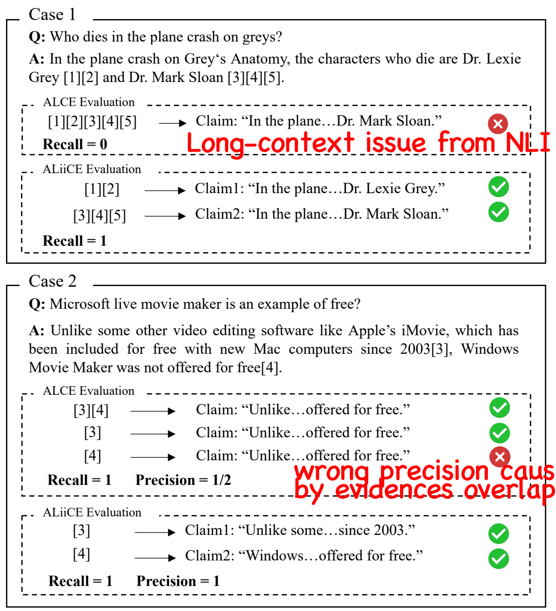


Figure 3: The results of sentence-level evaluation and ALiiCE show differences on the ASQA dataset.

Case Study



Case 1

Q: Who dies in the plane crash on greys?

A: In the plane crash on Grey‘s Anatomy, the characters who die are Dr. Lexie Grey [1][2] and Dr. Mark Sloan [3][4][5].

ALCE Evaluation

[1][2][3][4][5] → Claim: “In the plane…Dr. Mark Sloan.” ✗

Recall = 0

Long-context issue from NLI

ALiiCE Evaluation

[1][2] → Claim1: “In the plane…Dr. Lexie Grey.” ✓

[3][4][5] → Claim2: “In the plane…Dr. Mark Sloan.” ✓

Recall = 1

Case 2

Q: Microsoft live movie maker is an example of free?

A: Unlike some other video editing software like Apple’s iMovie, which has been included for free with new Mac computers since 2003[3], Windows Movie Maker was not offered for free[4].

ALCE Evaluation

[3][4] → Claim: “Unlike…offered for free.” ✓

[3] → Claim: “Unlike…offered for free.” ✓

[4] → Claim: “Unlike…offered for free.” ✗

Recall = 1 **Precision = 1/2**

wrong precision caused by evidences overlap

ALiiCE Evaluation

[3] → Claim1: “Unlike some…since 2003.” ✓

[4] → Claim2: “Windows…offered for free.” ✓

Recall = 1 **Precision = 1**

Figure 4: Case study.

More Insights

- Human evaluation shows strong alignment with ALiiCE.
- ALiiCE has a higher decision threshold.
- Open-source LLMs display great progress.
- Current methods on citation evaluation ignore the judgment of citation utility.