SelfCite: Self-Supervised Alignment for Context Attribution in Large Language Models

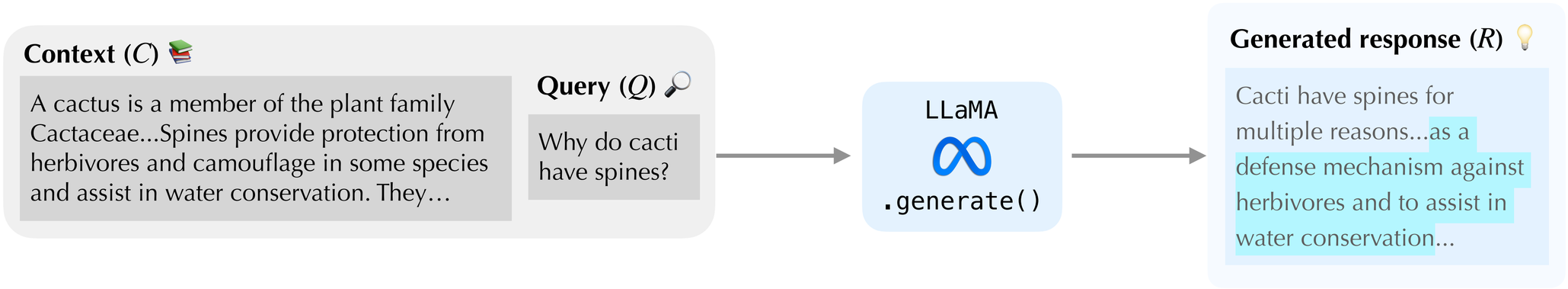
Yung-Sung Chuang[1], Ben Cohen-Wang[1], Shannon Shen[1], Zhaofeng Wu[1], Hu Xu[2], Xi Victoria Lin[2], James Glass[1], Shang-Wen Li[2], Wen-tau Yih[2]

MIT CSAIL[1], Meta FAIR[2]

ICML 2025

https://arxiv.org/abs/2502.09604

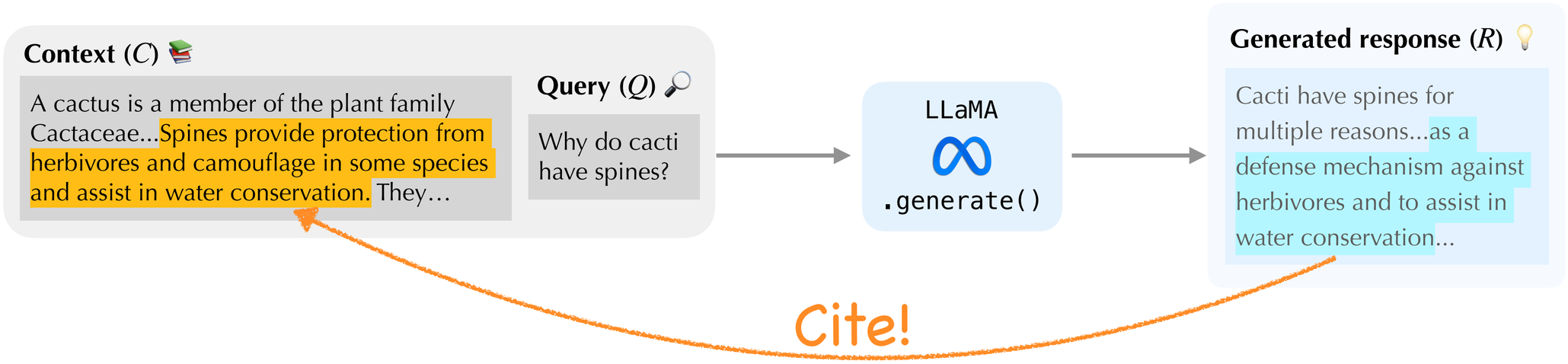
LLMs might hallucinate, so we need "citations"



Even if the LLM only hallucinates at 1% of the time,
the users will need to verify the content every time.

The verification process takes a lot of time from users,
especially for reasoning on long-context documents…

LLMs might hallucinate, so we need "citations"



Sentence-level citations allow the users to easily verify the generated content!

Anthropic's solution

Claude Citations API:

Sentence-level citations

Claude API Solutions Research Commitments Learn News Try Claude

Introducing Citations on the Anthropic API

Jan 23, 2025 • 3 min read

Today, we're launching Citations, a new API feature that lets Claude ground its answers in source documents. Claude can now provide detailed references to the exact sentences and passages it uses to generate responses, leading to more verifiable, trustworthy outputs.

Citations is generally available on the Anthropic API and Google Cloud's

Open-source solutions

LongCite: Data Generation from Proprietary APIs for SFT (Zhang et al., 2024)

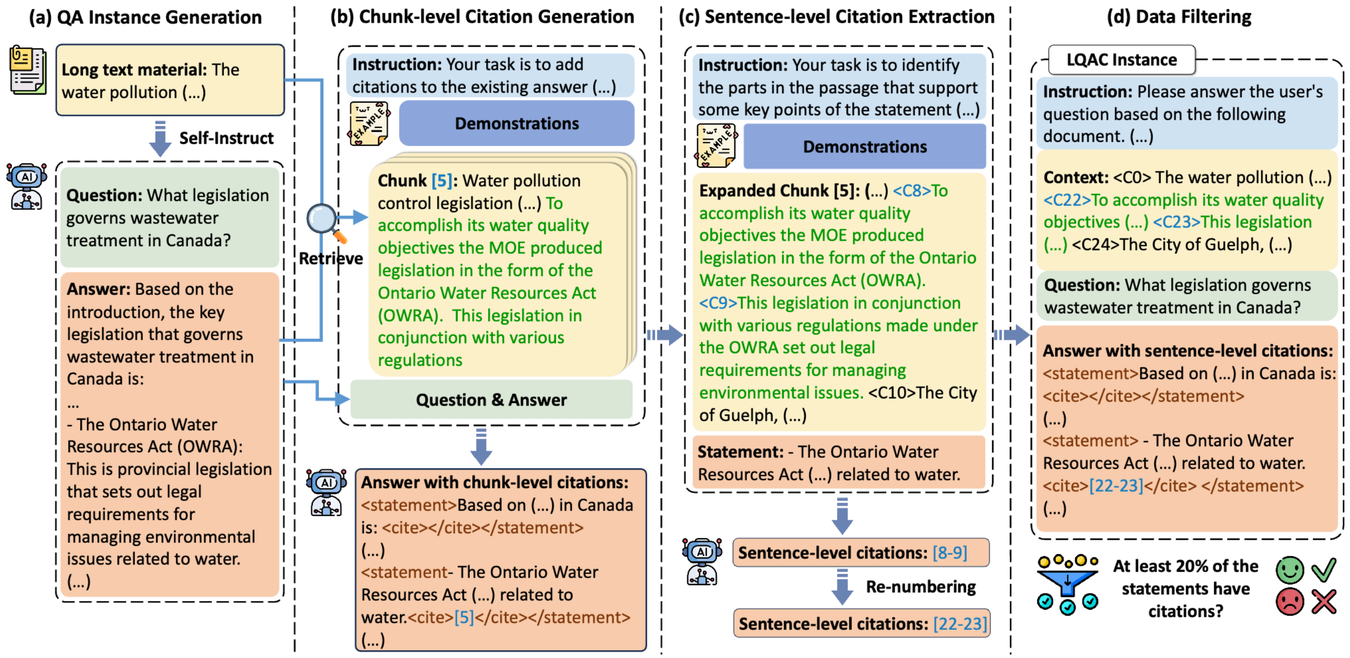
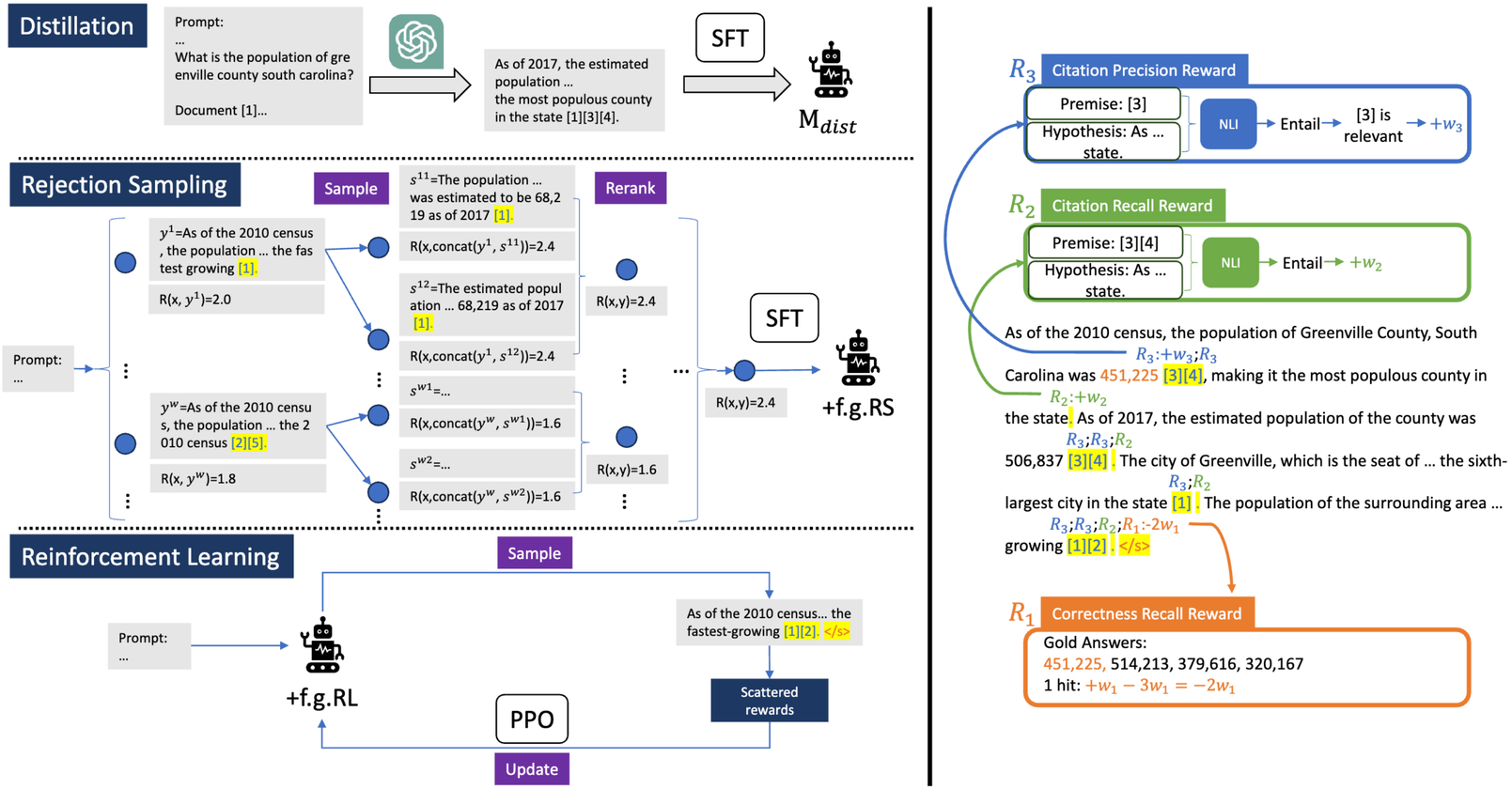


Figure 2: Overview of our CoF pipeline. The pipeline consists of four steps: (1) Generating long-context QA instance via Self-Instruct; (2) Using the answer to retrieve k context chunks and generating chunk-level citations; (3) Extracting sentence-level citations for each statement from the cited chunks. (4) Filter out LQAC instances with few citations.

Zhang, J., Bai, Y., Lv, X., Gu, W., Liu, D., Zou, M., Cao, S., Hou, L., Dong, Y., Feng, L. and Li, J., 2024. LongCite: Enabling llms to generate fine-grained citations in long-context QA. *arXiv preprint arXiv:2409.02897*.

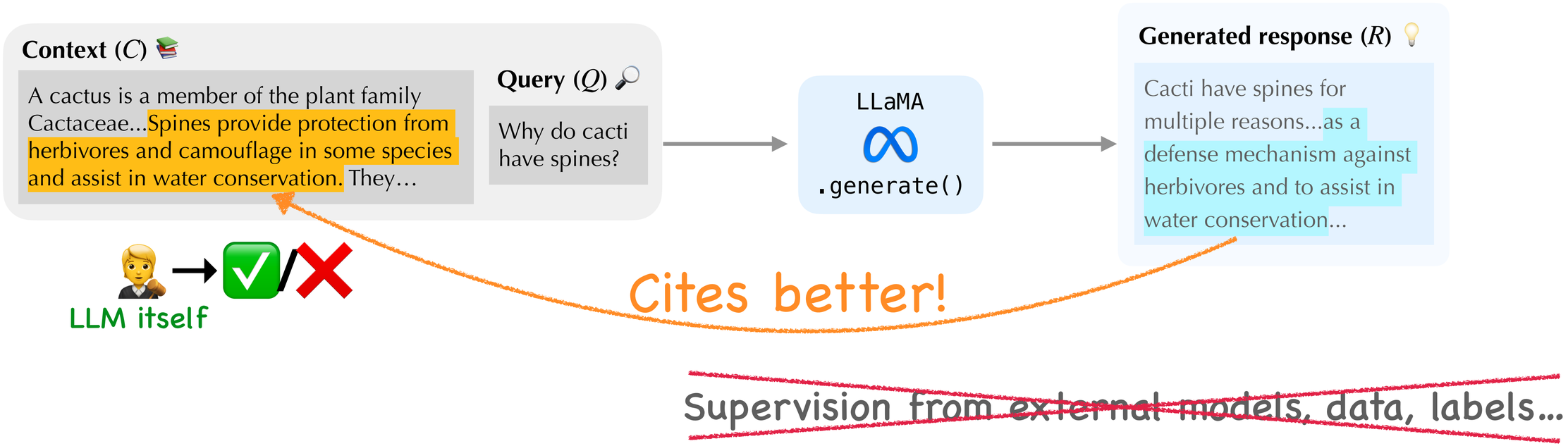
Open-source solutions

Supervision from external NLI models (Huang et al., 2024)



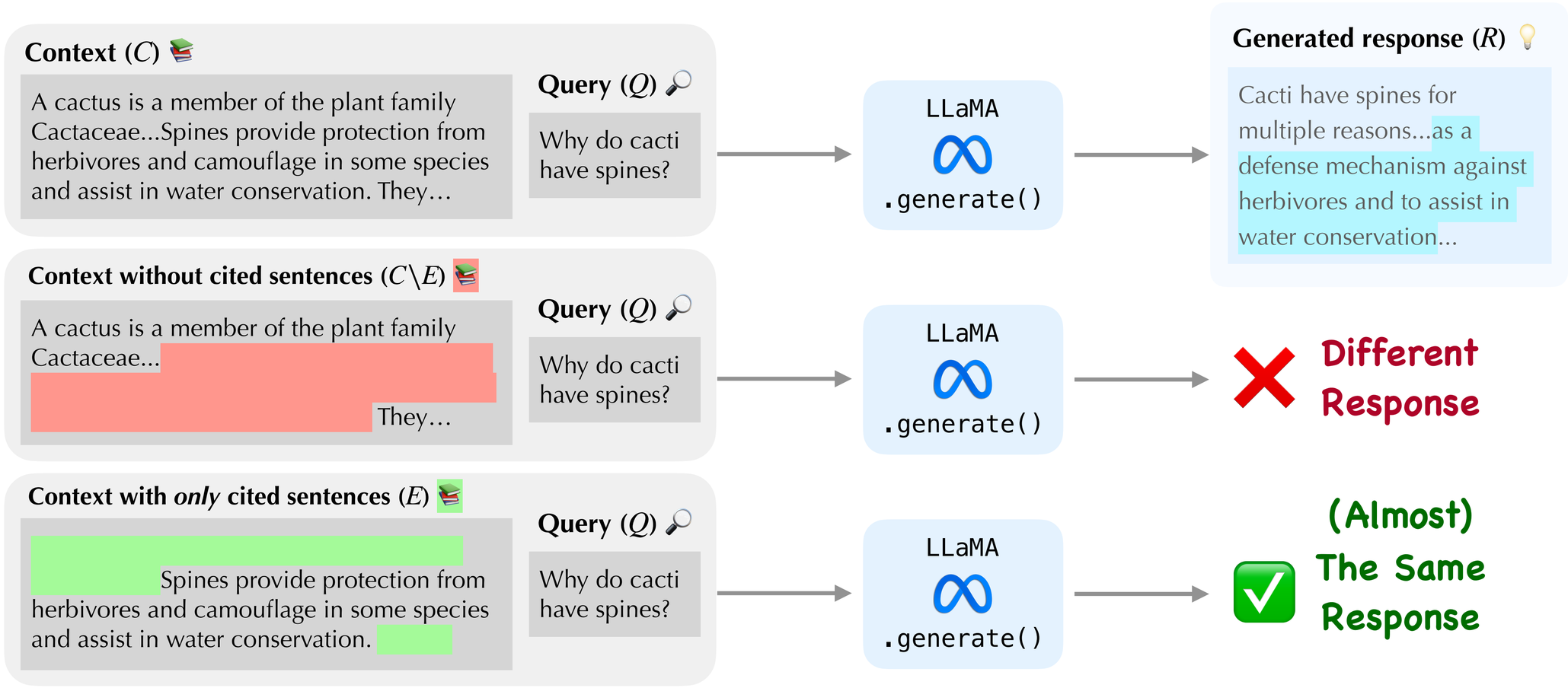
Huang, C., Wu, Z., Hu, Y. and Wang, W., 2024, August. Training Language Models to Generate Text with Citations via Fine-grained Rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2926-2949).

Can we make LLMs cites without external supervision?



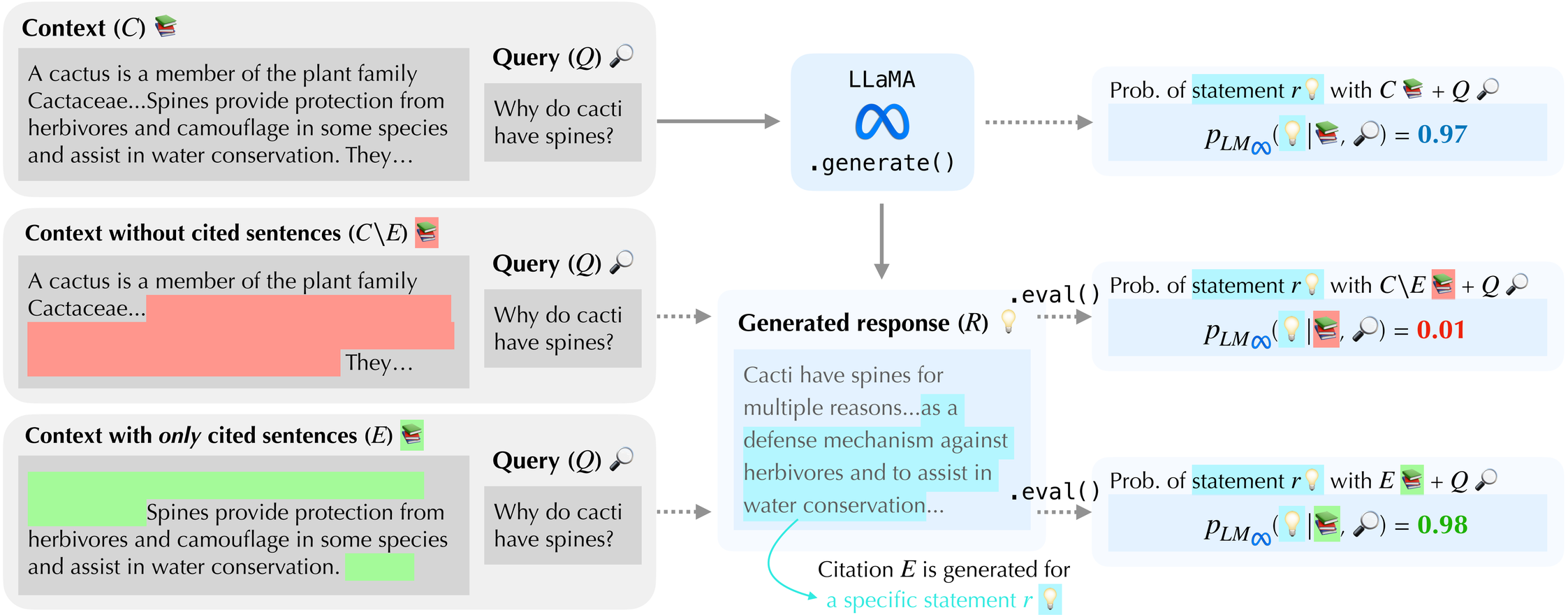
We build a Self-Supervised rewarding mechanism via 🔑 Context Ablation

Context Ablation



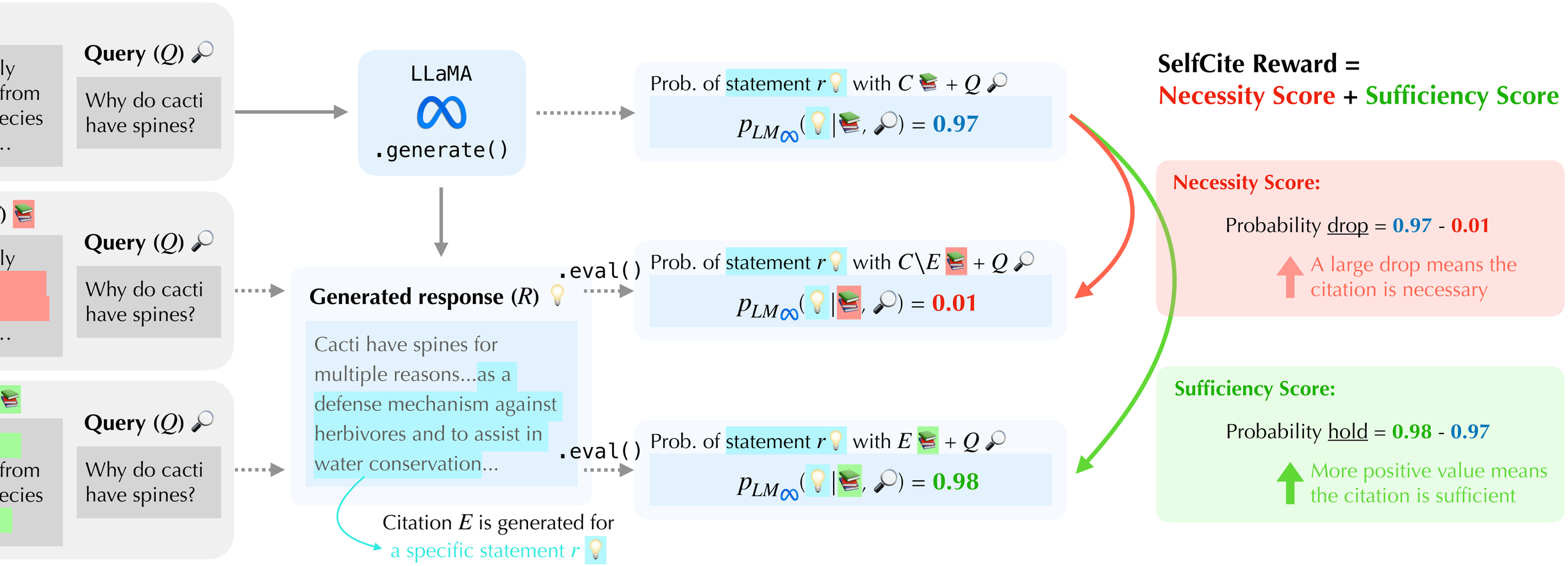
Context Ablation

Measure the log prob of generating the original response!



Context Ablation

Measure the log prob of generating the original response!



Reward Design

$$\text{Prob-Drop}(e_i) = \log p_{\text{LM}}(r_i \mid C) - \log p_{\text{LM}}(r_i \mid C \backslash E_i)$$

$$\text{Prob-Hold}(e_i) = \log p_{\text{LM}}(r_i \mid E_i) - \log p_{\text{LM}}(r_i \mid C)$$

$$\text{Reward}(e_i) = \text{Prob-Drop}(e_i) + \text{Prob-Hold}(e_i)$$

$$= \log p_{\text{LM}}(r_i \mid E_i) - \log p_{\text{LM}}(r_i \mid C \backslash E_i)$$

→ Only two forward passes needed!

Reward Use #1: Best-of-N Sampling (BoN)

Resample the sentence numbers within the citation tags*, e.g., <cite>[3-5]</cite>

Algorithm 1 SelfCite Best-of-N Sampling for Citations

Require: LM p_{LM}, context C, query Q, response R, number of candidates N

for $r_i \in R$ **do**

for $k = 1, \ldots, N$ **do**

$e_i^{(k)} \sim p_{\text{LM}}(\cdot \mid r_i, C, Q)$

$\text{Reward}(e_i^{(k)}) = \log p_{\text{LM}}\left(r_i \mid E_i^{(n)}\right) - \log p_{\text{LM}}\left(r_i \mid C \setminus E_i^{(n)}\right)$

end for

$e_i^* = \arg\max_k \ \text{reward}\left(e_i^{(k)}\right)$

end for

return $R^* = \{r_1, e_1^*, \ldots, r_S, e_S^*\}$

*We follow the format of LongCite (Zhang et al. 2024) to generate html tags to cite sentence numbers from the documents

Reward Use #2: SimPO Fine-tuning

- Use the long context document + query (but not answers!) from LongCite-45k data
- Use the BoN results (N=10) to create preference pairs
- Use SimPO[1], a variant of DPO but without a reference model, to save GPU memory

[1] Meng, Y., Xia, M., & Chen, D. (2024). SimPO: Simple preference optimization with a reference-free reward. Advances in Neural Information Processing Systems, 37, 124198-124235.

Experiments

- Start from LongCite-8B:
- BoN and SimPO both improve ~4 points on Citation F1
- SimPO + BoN improves 5.3 points on Citation F1

Model	Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport			Avg.	Citation
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	**F1**	**Length**
Ours: SelfCite																	
LongCite-8B (Our repro.)	67.0	78.1	66.6	74.8	90.7	79.9	60.8	77.9	64.1	67.1	87.2	73.7	81.6	89.3	84.5	73.8	83.5
+ BoN	68.4	81.3	71.2	76.1	92.8	81.2	67.2	81.0	68.8	70.6	90.9	76.9	**87.6**	92.4	**89.3**	77.5	93.4
+ SimPO	68.1	79.5	69.1	75.5	92.6	81.0	**69.4**	82.3	**71.5**	72.7	91.6	78.9	86.4	92.9	89.1	77.9	105.7
+ SimPO then BoN	**73.3**	79.4	**72.8**	76.7	**93.2**	82.2	**69.4**	**83.0**	71.1	**74.2**	**92.2**	**80.3**	86.7	92.7	89.2	**79.1**	94.7

Experiments

- Better than prompting proprietary API models
- Our 8B model best result (SimPO + BoN) is only behind Claude Citations by 2%

Model	Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport			Avg.	Citation
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	**F1**	**Length**
Proprietary models																	
GPT-4o†	46.7	53.5	46.7	**79.0**	87.9	80.6	55.7	62.3	53.4	65.6	74.2	67.4	73.4	90.4	79.8	65.6	220
Claude-3-sonnet†	52.0	67.8	55.1	64.7	85.8	71.3	46.4	65.8	49.9	67.7	89.2	75.5	77.4	93.9	84.1	67.2	132
GLM-4†	47.6	53.9	47.1	72.3	80.1	73.6	47.0	50.1	44.4	73.4	82.3	75.0	82.8	93.4	87.1	65.4	169
Ours: SelfCite																	
LongCite-8B (Our repro.)	67.0	78.1	66.6	74.8	90.7	79.9	60.8	77.9	64.1	67.1	87.2	73.7	81.6	89.3	84.5	73.8	83.5
+ BoN	68.4	81.3	71.2	76.1	92.8	81.2	67.2	81.0	68.8	70.6	90.9	76.9	**87.6**	92.4	**89.3**	77.5	93.4
+ SimPO	68.1	79.5	69.1	75.5	92.6	81.0	**69.4**	82.3	71.5	72.7	91.6	78.9	86.4	92.9	89.1	77.9	105.7
+ SimPO then BoN	**73.3**	79.4	**72.8**	76.7	93.2	82.2	**69.4**	83.0	71.1	74.2	92.2	80.3	86.7	92.7	89.2	79.1	94.7
Topline																	
Claude Citations	61.2	**81.7**	67.8	76.8	**98.4**	**84.9**	61.9	**94.1**	**72.9**	**88.5**	**99.7**	**93.2**	79.4	**99.2**	87.7	**81.3**	88.8

Thank you!

Code & Model: https://github.com/facebookresearch/SelfCite

Blog Post: https://voidism.github.io/SelfCite/

Paper: https://arxiv.org/abs/2502.09604

Yung-Sung Chuang
Email: *yungsung@mit.edu*