School of Computing
UNIVERSITY OF GEORGIA

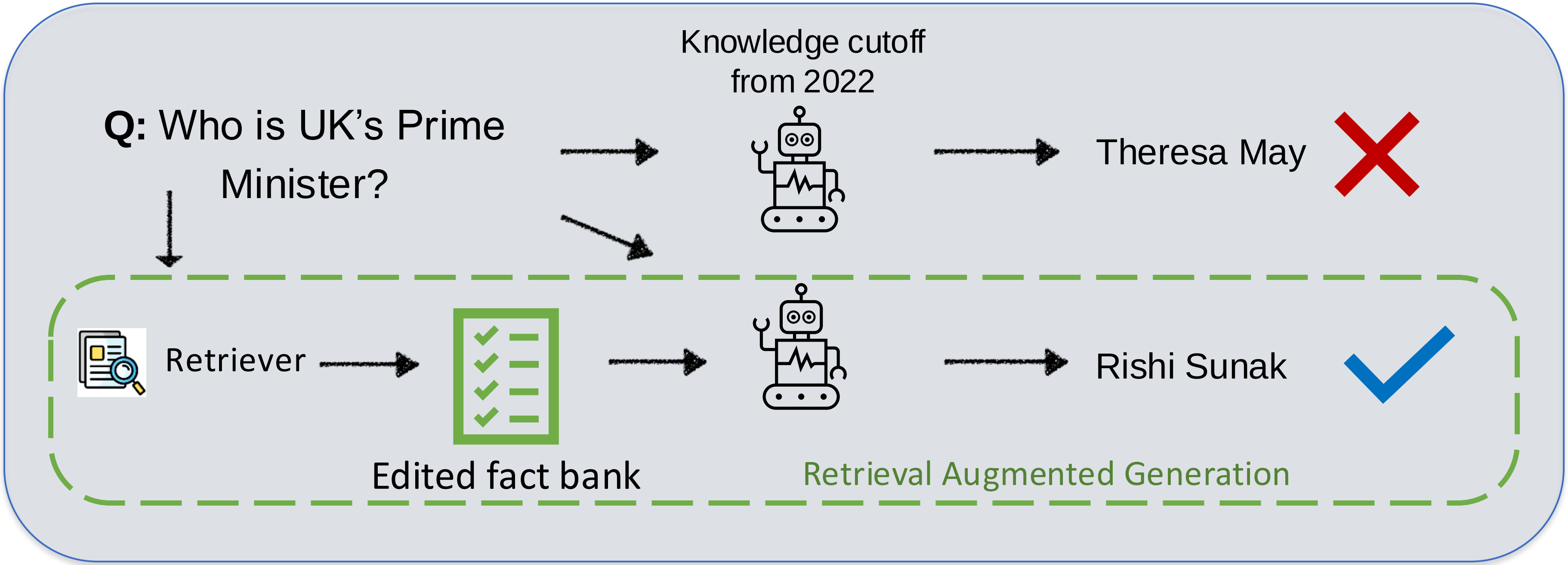
Retrieval-enhanced Knowledge Editing in Language Models for Multi-Hop Question Answering

Yucheng Shi[1], Qiaoyu Tan[2], Xuansheng Wu[1], Shaochen Zhong[3], Kaixiong Zhou[4], Ninghao Liu[1]

1. School of Computing, University of Georgia
2. Computer Science Department, NYU Shanghai
3. Computer Science Department, Rice University
4. Department of Electrical and Computer Engineering, NCSU

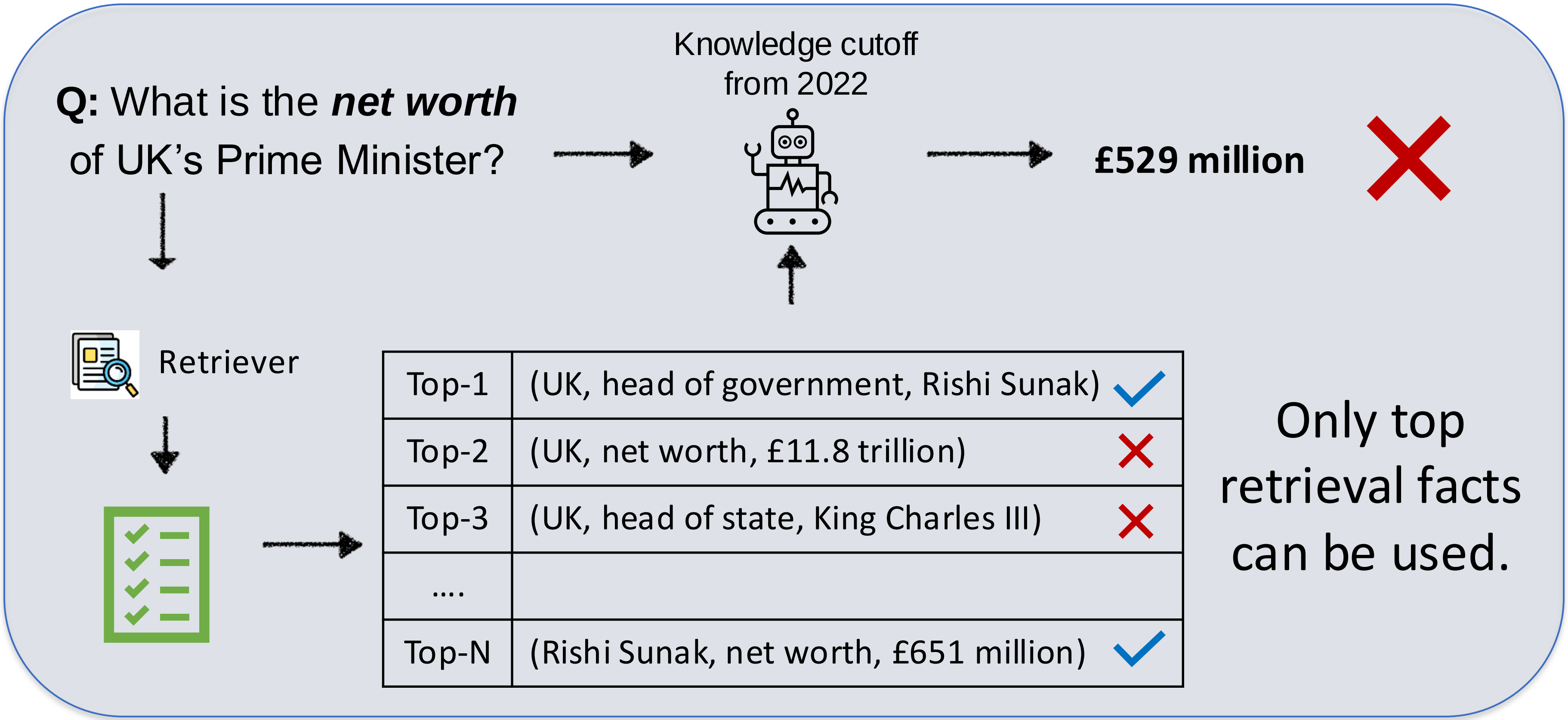
1. Background

Model Editing (Knowledge Editing) aims to edit LLMs to answer questions with updated knowledge.



Retrieval Augmented Generation (RAG) seem a good solution.
So, what is the *problem*?

1. Background- Editing for Multi-hop questions



Top-1	(UK, head of government, Rishi Sunak) ✓
Top-2	(UK, net worth, £11.8 trillion) ✗
Top-3	(UK, head of state, King Charles III) ✗
….	
Top-N	(Rishi Sunak, net worth, £651 million) ✓

Problem:

Facts beyond one-hop is ***hard*** to retrieve by current retrieval method (e.g., BM25 or DPR).

2. Retrieval-augmented Knowledge Editing (RAE)

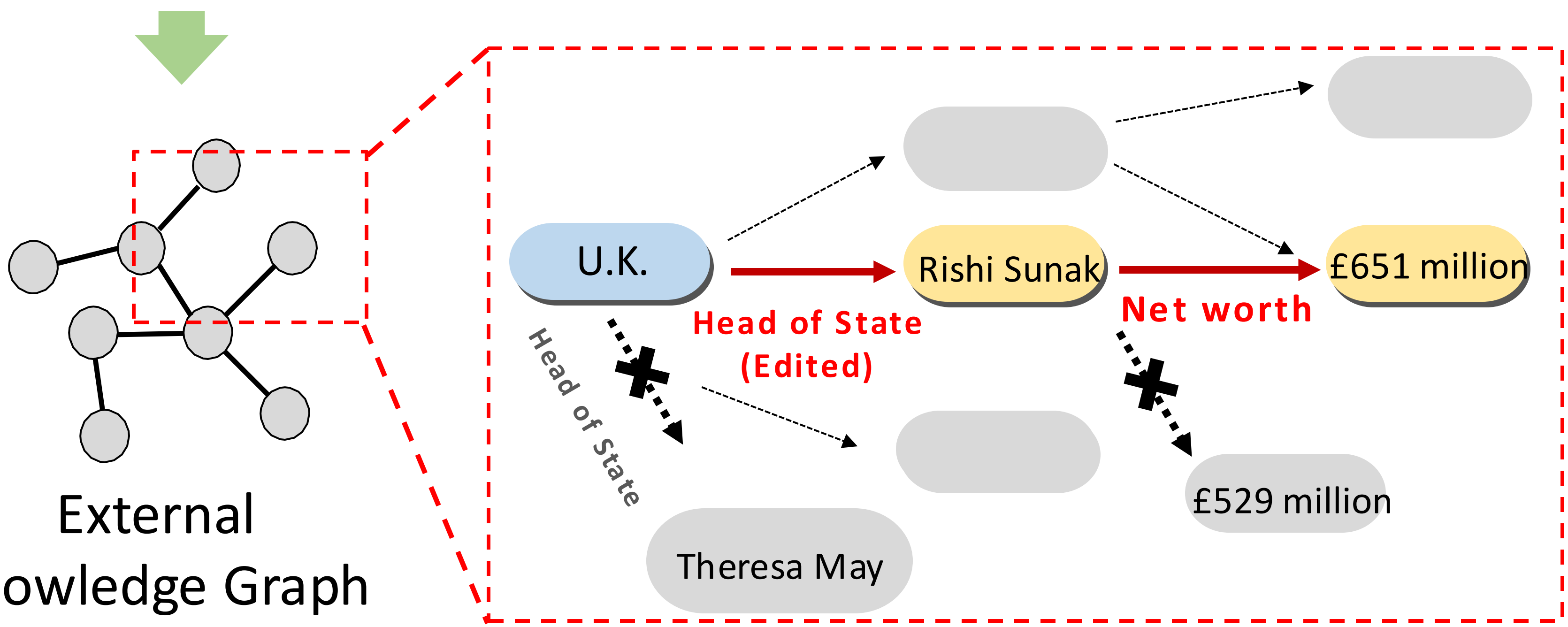
Step 1: Multi-hop Knowledge Retrieval

Edited fact

(UK, head of government, **Rishi Sunak**)

Original fact

(UK, head of government, **Theresa May**)

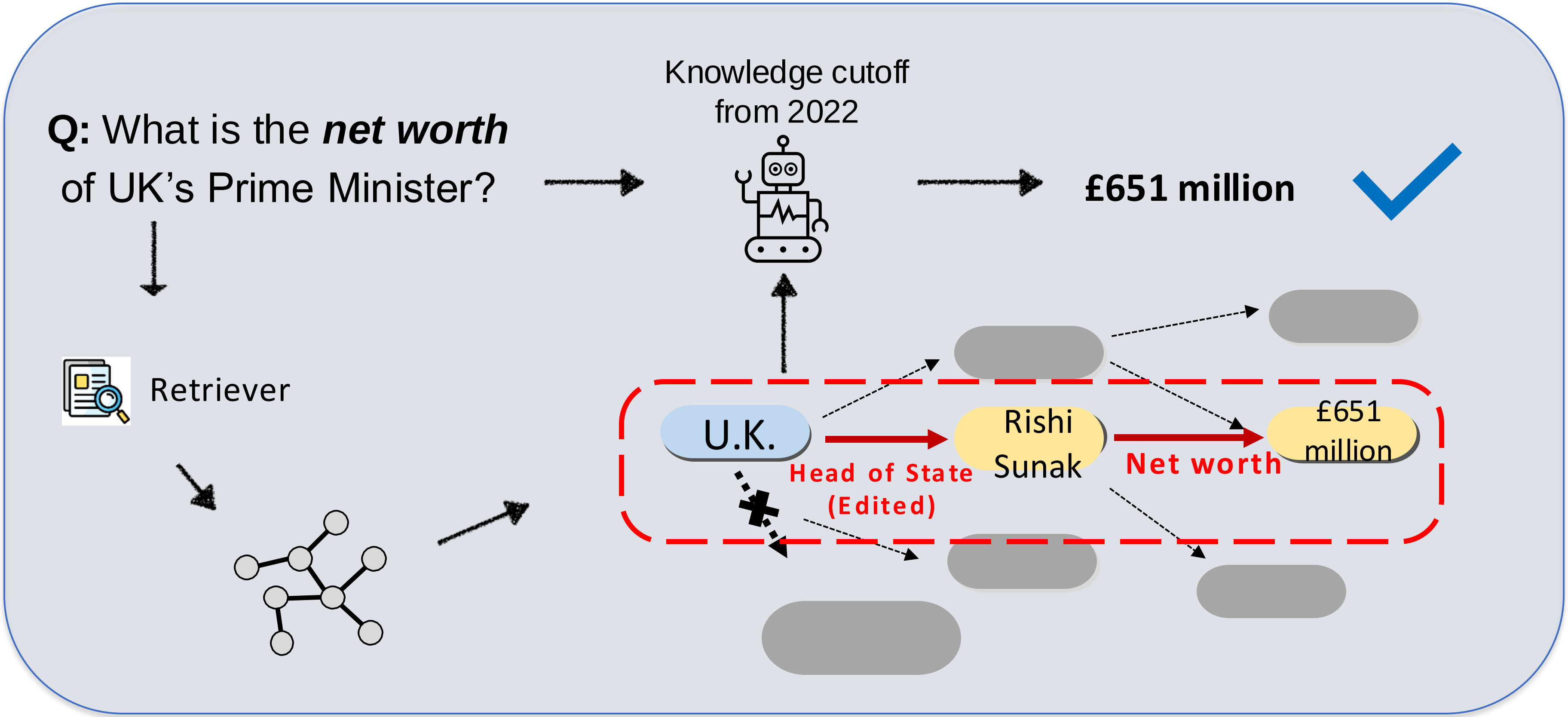


External Knowledge Graph (i.e., Wikidata5m)

Q: What is the ***net worth*** of UK's Prime Minister?

2. Retrieval-augmented Knowledge Editing (RAE)

Step 2: In-context learning for editing



3. MI-Based Retrieval Objective

Retrieval Objective

Maximize mutual information between subgraph and questions:

$$\max_{G_S} I\,(Q;\,G_S) \;=\; H(Q) \;-\; H(Q \mid G = G_S)$$

where Q are questions whose answers require editing,
G_S is our retrieval knowledge graph,
I denotes mutual information, H denotes entropy.

To simplify the setting, we consider one multi-hop question q at a time, which can be reformulated as:

$$\max_{G_S} \frac{p(q, G = G_S)}{p(G = G_S)} \log_2 \frac{p(q, G = G_S)}{p(G = G_S)}$$

4. Next Fact Prediction

Retrieved facts G_S are connected triplets:

$$G_S = (h_1, r_1, t_1, \dots, h_n, r_n, t_n)$$

where h, r, t are the head entity, relation, and tail entity.

Probabilities Estimation

Probabilities decomposition by Conditional Probability.

$$\frac{p(q, G = G_S)}{p(G = G_S)} = \frac{p(r_1, t_1, h_2, r_2, t_2, \dots, h_n, r_n, t_n | q, h_1)}{p\,(r_1, t_1, h_2, r_2, t_2, \dots, h_n, r_n, t_n | h_1)} \cdot \boxed{\frac{p(q, h_1)}{p(h_1)}}$$

$\boxed{\frac{p(q, h_1)}{p(h_1)}}$: Fixed value when given a specific question q

4. Next Fact Prediction

Probabilities Estimation

Probabilities decomposition by Conditional Probability.

Prob_A

$$\frac{p(q, G = G_S)}{p(G = G_S)} = \frac{\boxed{p(r_1, t_1, h_2, r_2, t_2, \ldots, h_n, r_n, t_n | q, h_1)}}{p\,(r_1, t_1, h_2, r_2, t_2, \ldots, h_n, r_n, t_n | h_1)} \cdot \frac{p(q, h_1)}{p(h_1)}$$

First, ***Prob_A*** has a recursive pattern:

Prob_B

$$\boxed{p(r_1, t_1, h_2, r_2, t_2, \ldots, h_n, r_n, t_n | q, h_1)}$$
$$= p(\, t_1, h_2, r_2, t_2, \ldots, h_n, r_n, t_n \,| q, h_1, r_1) \cdot \underline{p\,(r_1 \,|\, q, h_1)}$$

We can solve it step by step. But how to estimate ***Prob_B*** ?

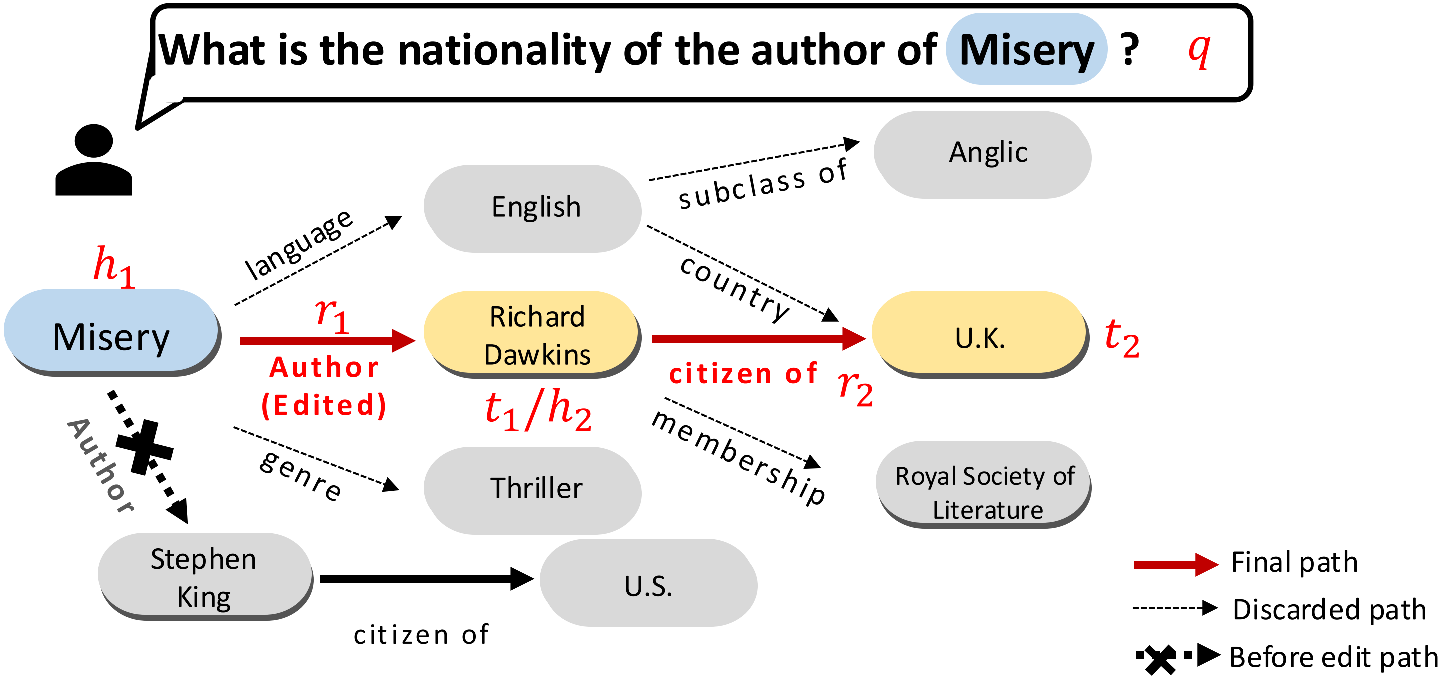
4. Next Fact Prediction

Next Fact Prediction

The probability $p\ (r_1\ |q, h_1)$ for each candidate relation can be estimated by an auto-regressive language model.

Prob$_B$

$$p(\,r_1\,|\,q,h_1) \approx \prod_{i=1}^{|r_1|} f_\phi(\,w_{r_1}^{(i)}|w_q^{(1)},\ldots,w_q^{|q|},w_{h_1}^{(1)},\ldots,w_{h_1}^{(|h_1|)},w_{r_1}^{(1)},\ldots,w_{r_1}^{(i-1)})$$



5. Redundant Knowledge Pruning

- **Irrelevant facts** can mislead the LLM.
- We propose to **Prune** facts to avoid hallucinations.

How?

Use the LLM's **output entropy** as an indicator of uncertainty.

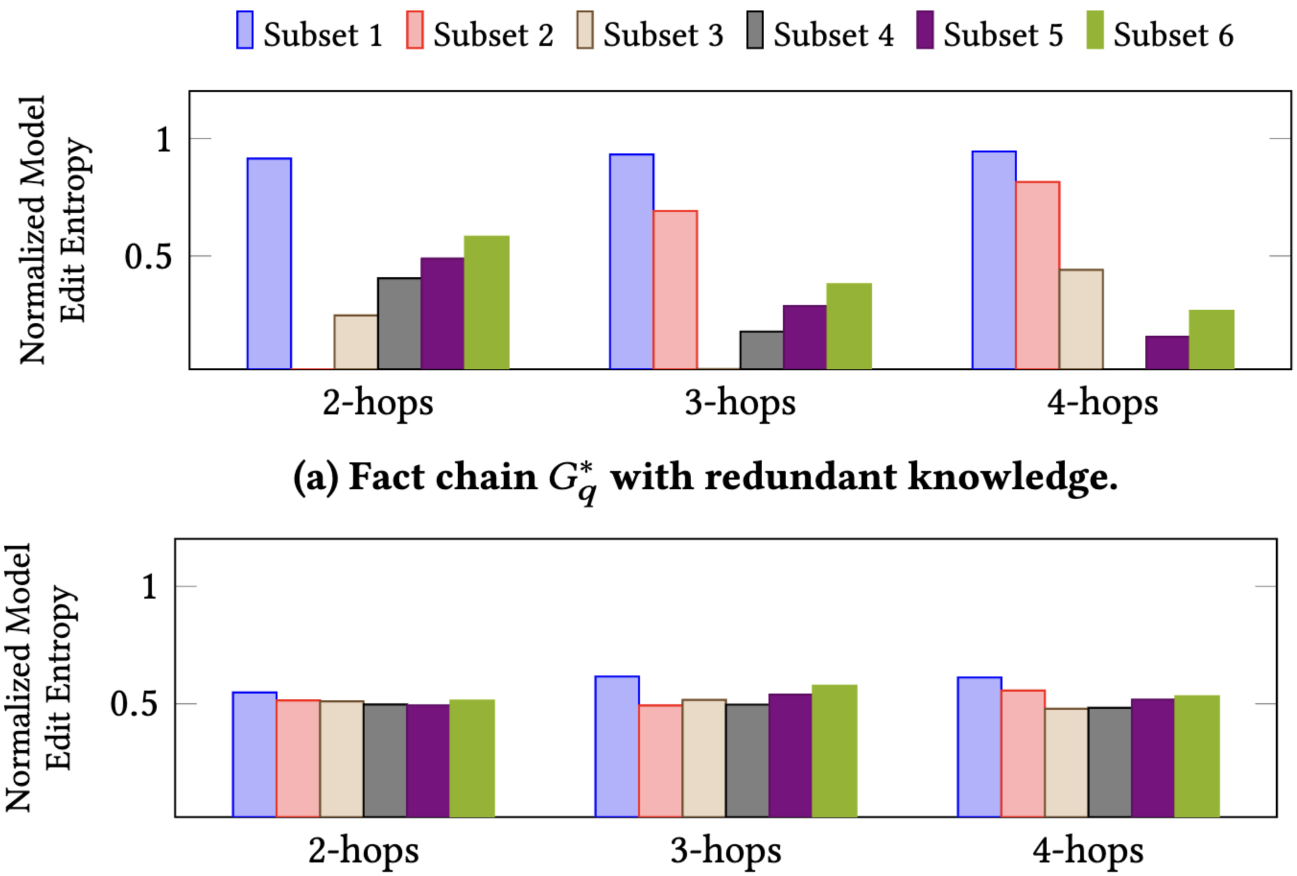
Editing Uncertainty

We define it as the entropy of the output generated by large language models.

$$H(Y \mid X = x) = -\sum_{y} p(y \mid x) \log_2 p(y|x)$$

- Lower entropy indicates more confidence.
- We use ***facts with lowest entropy*** as retrieval results.

5. Redundant knowledge Pruning



(a) Fact chain G_q^* with redundant knowledge.

(b) Random fact chain without useful knowledge.

The entropy is ***minimized*** when the retrieved facts are precisely those required to answer the question.

6. Experiments - Settings

To evaluate editing performance, we use three multi-hop edit datasets:

- **MQUAKE-CF** (3000 cases)
- **MQUAKE-T** (1868 cases)
- **Popular** (274 cases)

We compare with six baseline methods, including:

- **Model weight updating methods:**
  - *Fine Tune, ROME, MEMIT*
- **Auxiliary models methods:**
  - *SEARC*
- **In-context learning methods:**
  - *Mello, DeepEdit*

6. Experiments – Editing Performance

Language Models	Datasets	Fine Tune	ROME	MEMIT	SEARC	Mello	DeepEdit	Subgraph Retriever	RAE(ours)
		Editing Methods							
GPT-2 (1.5B)	M-CF	3.8	1.7	2.3	4.0	0.0	0.0	21.9	**62.8**
	M-T	5.8	6.4	1.6	2.7	0.0	0.0	20.3	**61.8**
	Popular	6.2	4.3	2.9	1.1	0.0	0.0	26.7	**47.1**
GPT-J (6B)	M-CF	7.7	7.6	8.1	6.8	15.3	9.3	36.2	**69.3**
	M-T	3.1	4.1	10.6	2.8	36.7	19.6	51.2	**63.9**
	Popular	6.8	7.5	4.4	1.3	12.8	6.6	45.8	**49.6**
Falcon (7B)	M-CF	5.6	1.7	2.3	7.9	10.7	10.8	40.1	**66.8**
	M-T	17.2	7.3	1.6	4.5	51.5	31.7	56.1	**61.6**
	Popular	2.1	4.0	1.1	3.0	8.1	9.5	43.0	**50.0**
Vicuna (7B)	M-CF	4.8	8.4	7.6	7.9	10.2	11.4	39.4	**67.2**
	M-T	23.1	5.0	1.7	4.5	51.7	40.4	58.6	**63.2**
	Popular	4.0	3.8	2.4	3.0	7.7	8.2	29.5	**36.1**
Llama2 (chat) (7B)	M-CF	5.4	6.3	3.8	7.9	20.7	11.2	45.7	**69.1**
	M-T	17.1	8.7	1.7	4.5	49.4	37.9	63.1	**66.2**
	Popular	5.2	13.8	4.9	3.0	13.5	11.1	41.9	**51.4**

The multi-hop edited accuracy metrics is reported: if the edited answer appears in the final output, it is a correct edit.

7. Experiments – Multi-hop Fact Retrieval Performance

MQUAKE-CF							
Question Type		2-hops		3-hops		4-hops	
Category	Retrieval	P@1	P@2	P@1	P@3	P@1	P@4
Embedding	KG Link	52.7	28.7	18.2	3.7	14.0	0.0
	QR	62.3	7.7	14.7	0.0	12.3	0.0
	Mello(Llama2)	84.3	80.0	80.7	42.3	83.3	25.7
Probability	SR(GPT-2)	77.7	50.3	67.3	25.3	65.0	20.0
	SR(Llama2)	78.3	55.7	79.7	37.0	69.3	28.7
Mutual Information	RAE(GPT-2)	83.0	66.3	77.3	41.0	80.3	43.7
	RAE(GPT-J)	83.0	69.7	81.3	53.7	82.7	54.0
	RAE(Falcon)	82.3	70.7	72.3	44.3	81.7	47.3
	RAE(Vicuna)	81.0	66.7	79.3	50.3	85.0	50.0
	RAE(Llama2)	82.7	69.3	84.0	49.3	82.0	47.0

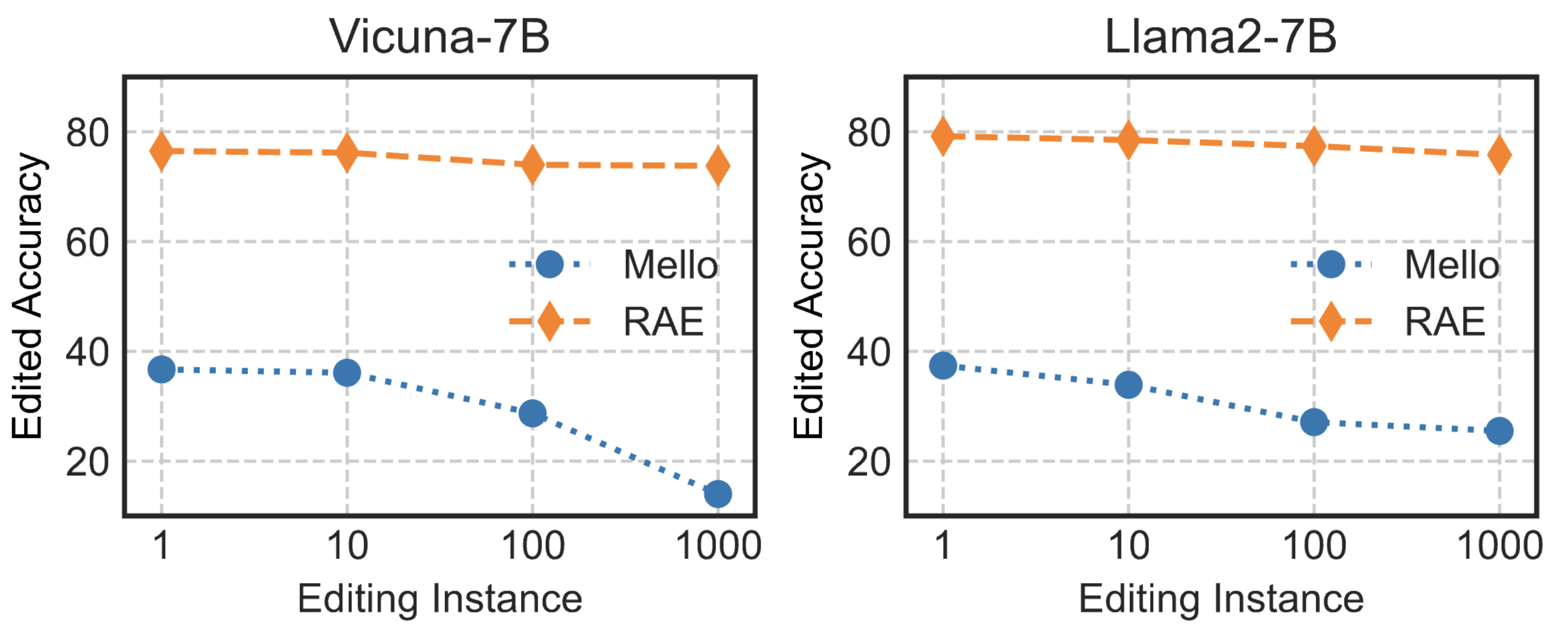
We use the metric $Precision@K$, which calculates the proportion of relevant facts within the top K results: $Precision@K$ = |{relevant facts}|/K × 100%, abbreviated as $P@K$.

7. Experiments – Pruning improves editing performance

Dataset	MQUAKE-CF					
Type	Strategy	GPT-2	GPT-J	Falcon	Vicuna	Llama2 (chat)
2-hops	w/o Pruning	63.0	63.7	65.2	63.8	70.1
	w/ Pruning	73.3	75.5	74.5	73.5	75.8
	Gain	16.3%↑	18.5%↑	14.3%↑	15.2%↑	8.1%↑
3-hops	w/o Pruning	43.1	53.8	55.6	55.0	60.3
	w/ Pruning	53.2	65.4	62.1	62.7	65.8
	Gain	23.4%↑	21.6%↑	11.7%↑	14.0%↑	9.1%↑
4-hops	w/o Pruning	49.9	58.8	55.2	61.5	61.6
	w/ Pruning	61.9	66.9	62.9	65.5	65.8
	Gain	24.0%↑	13.8%↑	13.9%↑	6.5%↑	6.8%↑

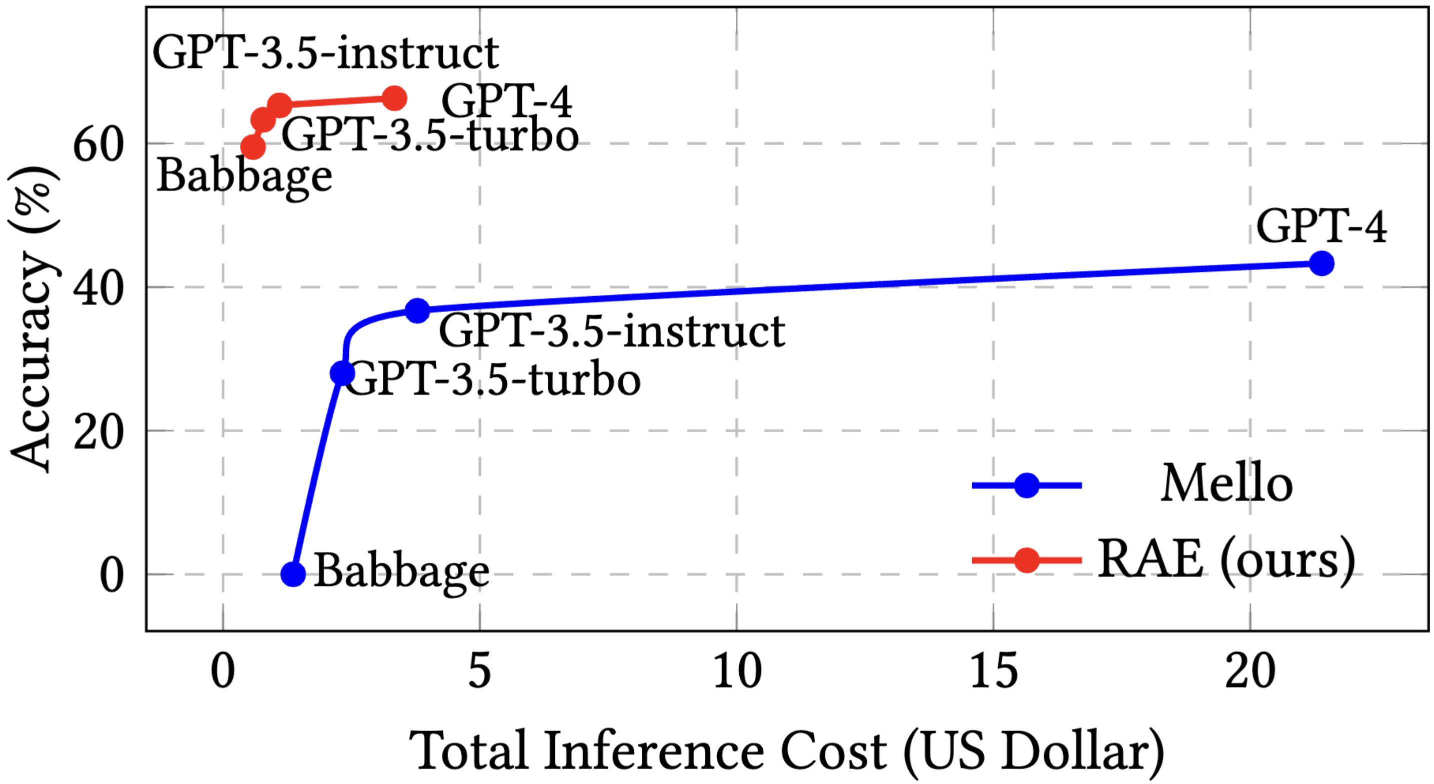
Pruning achieves an average accuracy improvement of **14.5%** across various language models.

7. Experiments – Performance with batch size



RAE's accuracy remains **stable** with increasing editing instances, whereas Mello's accuracy significantly **declines** with increasing instances.

7. Experiments – RAE works with proprietary models



RAE achieves **better** editing performance with **lower** inference cost over different proprietary models.

Acknowledgements

- DATA Lab @UGA and collaborators

DATA Lab @ UGA

UNIVERSITY OF GEORGIA

上海纽约大学 NYU SHANGHAI

NC STATE UNIVERSITY

RICE

- Funding agencies:
  - National Science Foundation.

- Everyone attending the talk!

Q & A